

1 **A systematic comparison reveals substantial differences in chromosomal versus episomal**
2 **encoding of enhancer activity**

3 ***Running title: Comparing chromosomal and episomal reporter assays***

4
5 Fumitaka Inoue¹†, Martin Kircher²†, Beth Martin², Gregory M. Cooper³, Daniela M. Witten⁴,
6 Michael T. McManus⁵, Nadav Ahituv^{1*}, Jay Shendure^{2,6*}

7
8 **Affiliations:**

9 ¹Department of Bioengineering and Therapeutic Sciences, Institute for Human Genetics,
10 University of California San Francisco, San Francisco CA, USA

11 ²Department of Genome Sciences, University of Washington, Seattle WA, USA

12 ³HudsonAlpha Institute for Biotechnology, Huntsville AL, USA.

13 ⁴Departments of Statistics and Biostatistics, University of Washington, Seattle WA, USA

14 ⁵Department of Microbiology and Immunology, UCSF Diabetes Center, Keck Center for
15 Noncoding RNA, University of California, San Francisco, San Francisco, CA, USA

16 ⁶Howard Hughes Medical Institute, Seattle WA, USA

17

18 †These authors contributed equally to this work

19 *Correspondence to nadav.ahituv@ucsf.edu and shendure@uw.edu

20

21

22 **Abstract**

23

24 Candidate enhancers can be identified on the basis of chromatin modifications, the binding of
25 chromatin modifiers and transcription factors and cofactors, or chromatin accessibility. However,
26 validating such candidates as bona fide enhancers requires functional characterization, typically
27 achieved through reporter assays that test whether a sequence can drive expression of a
28 transcriptional reporter via a minimal promoter. A longstanding concern is that reporter assays
29 are mainly implemented on episomes, which are thought to lack physiological chromatin.
30 However, the magnitude and determinants of differences in *cis*-regulation for regulatory
31 sequences residing in episomes versus chromosomes remain almost completely unknown. To
32 address this question in a systematic manner, we developed and applied a novel lentivirus-based
33 massively parallel reporter assay (lentiMPRA) to directly compare the functional activities of
34 2,236 candidate liver enhancers in an episomal versus a chromosomally integrated context. We
35 find that the activities of chromosomally integrated sequences are substantially different from the
36 activities of the identical sequences assayed on episomes, and furthermore are correlated with
37 different subsets of ENCODE annotations. The results of chromosomally-based reporter assays
38 are also more reproducible and more strongly predictable by both ENCODE annotations and
39 sequence-based models. With a linear model that combines chromatin annotations and sequence
40 information, we achieve a Pearson's R^2 of 0.347 for predicting the results of chromosomally
41 integrated reporter assays. This level of prediction is better than with either chromatin annotations
42 or sequence information alone and also outperforms predictive models of episomal assays. Our
43 results have broad implications for how *cis*-regulatory elements are identified, prioritized and
44 functionally validated.

45

46 Introduction

47
48 An enhancer is defined as a short region of DNA that can potentiate the expression of a gene,
49 independent of its orientation and flexible with respect to its position relative to the transcriptional
50 start site (Banerji et al. 1981; Moreau et al. 1981). Enhancers are thought to be modular, in the
51 sense that they are active in heterologous sequence contexts and in that multiple enhancers may
52 additively dictate the overall expression pattern of a gene (Shlyueva et al. 2014). They act through
53 the binding of transcription factors, which recruit histone modifying factors, such as histone
54 acetyltransferase (HAT) or histone methyltransferase (HMT). Enhancers are also associated with
55 chromatin remodeling factors (e.g. SWI/SNF) and the cohesin complex, which are involved in
56 regulating chromatin structure and accessibility (Schmidt et al. 2010; Euskirchen et al. 2011;
57 Faure et al. 2012).

58
59 Antibodies against specific transcription factors (TFs), histone modifications or transcriptional
60 co-activators are commonly used for chromatin immunoprecipitation followed by massively
61 parallel sequencing (ChIP-seq) to identify candidate enhancers in a genome-wide manner. For
62 example, the ENCODE Consortium and other efforts have identified thousands of candidate
63 enhancers in mammalian genomes on the basis of such marks or their correlates (e.g. p300 ChIP-
64 seq; H3K27ac ChIP-seq; DNaseI hypersensitivity) in diverse cell lines and tissues (Visel et al.
65 2009; Dunham I et al. 2012). However, a major limitation of such assays is that they reflect
66 biochemical marks that are correlated with enhancer activity, rather than directly showing that
67 any particular sequence actually functions as an enhancer. In other words, while such assays yield
68 genome-wide catalogs of potential enhancers, they do not definitively predict bona fide enhancers
69 nor precisely define their boundaries.

70
71 For decades, the primary means of functionally validating enhancers has been the episomal
72 reporter assay. The standard approach is to relocate the candidate enhancer sequence to an
73 episomal vector, adjacent to a minimal promoter driving expression of a reporter gene, e.g.
74 luciferase or others. More recently, massively parallel reporter assays (MPRAs) have enabled the
75 functional characterization of *cis*-regulatory elements including enhancers in a high-throughput
76 manner. MPRAs use sequencing-based quantification of reporter barcodes to enable multiplexing
77 of the reporter assay (Patwardhan et al. 2009). MPRAs have been used primarily in an episomal
78 manner for the saturation mutagenesis of promoters and enhancers (Patwardhan et al. 2009;
79 Kinney et al. 2010; Melnikov et al. 2012; Patwardhan et al. 2012), for exploring the grammatical
80 rules of promoters and enhancers (Smith et al. 2013; Sharon et al. 2014), and for testing of
81 thousands of enhancer candidates in different cells or tissues (Kwasnieski et al. 2012; Arnold et
82 al. 2013; Kheradpour et al. 2013; Arnold et al. 2014; Shlyueva et al. 2014; Savic et al. 2015;
83 White 2015). Adeno-associated virus (AAV) MPRAs have also been developed, allowing these
84 assays to be carried out *in vivo* and to perform reporter assays within target cells and tissues that

85 are difficult to transduce, such as the brain (Shen et al. 2016), although these do not involve
86 genomic integration.

87
88 Despite their widespread use to validate enhancers and other *cis*-regulatory elements, a
89 longstanding concern about reporter assays is that they are almost always carried out via transient
90 transfection of non-integrating episomes. It is unknown whether transiently transfected sequences
91 are chromatinized in a way that makes them appropriate models for endogenous gene expression
92 from chromosomes (Smith and Hager 1997); but to the extent that this question has been explored,
93 there are differences. For example, work from Archer, Hager and colleagues using the mouse
94 mammary tumor virus (*MMTV*) promoter as a model shows differences in histone H1
95 stoichiometry and nucleosome positioning resulting in an inability of episomal assays to reliably
96 assay cooperative TF binding (Archer et al. 1992; Smith and Hager 1997; Hebbar and Archer
97 2007; Hebbar and Archer 2008). In another work, the chromatin structure of transiently
98 transfected non-replicating plasmid DNA was observed to be differently fragmented than
99 endogenous chromatin by micrococcal nuclease, and along with other data supports a model in
100 which atypical chromatin might be induced by association of episomes with nuclear structures
101 (Jeong and Stein 1994). However, the extent to which these factors operate to confound the results
102 of enhancer reporter assays more broadly, for categorical (i.e. is a particular sequence an
103 enhancer?), qualitative (i.e. in what tissues is an element an enhancer?), and quantitative
104 validation (i.e. what level of activation does a particular sequence confer?), has yet to be
105 systematically investigated.

106
107 To address these questions, we developed lentiviral MPRA (lentiMPRA), a technology that uses
108 lentivirus to integrate enhancer MPRA libraries into the genome. To overcome the substantial
109 position-effect variegation observed by others in attempting to use lentiviral infection for MPRA
110 (Murtha et al. 2014), we employed a flanking antirepressor element (#40) and a scaffold-attached
111 region (SAR) (Klehr et al. 1991; Kwaks et al. 2003) on either side of our construct. In addition,
112 we relied on as many as 100 independent reporter barcode sequences per assayed candidate
113 enhancer sequence, integrated at diverse sites. The resulting system allows for high-throughput,
114 highly reproducible and quantitative measurement of the regulatory potential of candidate
115 enhancers in a chromosomally integrated context. Furthermore, the cell-type range of lentivirus
116 transduction is much broader than transfection, e.g. permitting MPRA to be conducted in
117 neurons, primary cells or organoids.

118
119 By using integration-competent vs. integration-defective components of the lentiviral system, we
120 directly compared the functional activities of 2,236 candidate liver enhancers in a chromosomally
121 integrated versus an episomal context in the human liver hepatocellular carcinoma cell line,
122 HepG2. We find that the activities of chromosomally integrated sequences are substantially
123 different from the activities of the identical sequences assayed episomally, and are correlated with
124 different subsets of ENCODE annotations. We also find that the results of chromosomally-based

125 reporter assays are more reproducible and more strongly predicted by ENCODE annotations and
126 sequence-based models.

127

128 **Results**

129

130 Construction and validation of the lentiMPRA vector

131

132 The potential for confounding of lentiviral assays by site-of-integration effects was demonstrated
133 by a recent MPRA study (Functional Identification of Regulatory Elements Within Accessible
134 Chromatin or FIREWACH) that used lentiviral infection and found that 26% of positive controls
135 did not show activated GFP expression, while other measures estimated a false positive rate of
136 22% (Murtha et al. 2014). We therefore constructed a lentiviral vector (pLS-mP) that contains a
137 minimal promoter (mP) and the enhanced green fluorescent protein (EGFP) gene flanked on one
138 side by the antirepressor element #40 and the other by a SAR that reduce site-of-integration effects
139 and provide consistent transgene expression (Fig. 1A, Supplementary File 1) (Klehr et al. 1991;
140 Kwaks et al. 2003; Kissler et al. 2006). In experiments involving chromosomal integration of this
141 enhancer reporter, we confirmed that EGFP is not expressed in the absence of an enhancer, while
142 abundantly expressed under the control SV40 enhancer across a panel of cell lines representing
143 diverse tissues-of-origin. These include: K562 (lymphoblasts), H1-ESC (embryonic stem cells),
144 HeLa-S3 (cervix), HepG2 (hepatocytes), T-47D (epithelial) and Sk-n-sh retinoic acid treated
145 (neuronal) cells (Fig. S1). Furthermore, when SV40 and the *Ltv1* liver enhancer (Patwardhan et al.
146 2012) are tested without the flanking antirepressor sequences, we observed much lower levels of
147 EGFP expression in HepG2 cells, consistent with our expectation that the antirepressors facilitate
148 robust enhancer-mediated expression from the integrated reporter (Fig. 1B).

149

150 Design and construction of a library of candidate liver enhancers

151

152 To evaluate lentiMPRA, we designed a liver enhancer library that comprises 2,236 candidate
153 sequences and 204 control sequences (Fig. 1C, Supplementary File 2), each 171 bp in length. All
154 enhancer candidate sequences were chosen on the basis of having ENCODE HepG2 ChIP-seq
155 peaks for EP300 and H3K27ac, which are generally indicative of enhancer function (Heintzman
156 et al. 2007; Visel et al. 2009). A subset of candidates (“type 1”) are centered at ChIP-seq peaks for
157 forkhead box A1 (FOXA1) or FOXA2, known liver pioneer transcription factors (Lupien et al.
158 2008) or hepatocyte nuclear factor 4 alpha (HNF4A), a nuclear receptor involved in lipid
159 metabolism and gluconeogenesis (Watt et al. 2003), while also overlapping with ENCODE-
160 derived ChIP-seq peaks for the cohesin complex (RAD21 and SMC3) or chromodomain helicase
161 DNA binding protein 2 (CHD2), a chromatin remodeler that is part of the SWI/SNF complex.
162 Other subsets of candidates were required to overlap only a liver transcription factor peak (“type
163 2”), only a chromatin remodeler peak (“type 3”), or neither (“type 4”). The 204 control sequences
164 comprised 200 synthetically designed controls from a previous study (synthetic regulatory element

165 sequences (SRESs); 100 positive & 100 negative) (Smith et al. 2013) and an additional 2 positive
166 (pos1 and pos2) and 2 negative endogenous controls (neg1 and neg2). We confirmed by standard
167 luciferase reporter assay that pos1 and pos2 showed weak and strong enhancer activity,
168 respectively, while neg1 and neg2 showed no activity (Fig. S2).

169
170 Each of the 2,440 enhancer candidates or controls was synthesized in *cis* with 100 unique reporter
171 barcodes on a 244,000-feature microarray (Agilent OLS; 15 bp primer + 171 bp enhancer
172 candidate or control + 14 bp spacer + 15 bp barcode + 15 bp primer = 230-mers). The purpose of
173 encoding a large number of barcodes per assayed sequence was to facilitate reproducible and
174 quantitative measurements of regulatory activity, as well as to mitigate against non-uniformity in
175 oligonucleotide synthesis. We cloned these oligonucleotides to a version of the lentiMPRA vector
176 that lacked mP and EGFP reporter. Subsequently, a restriction site in the spacer was used to reinsert
177 the mP + EGFP cassette between the candidate enhancer and barcode, thus positioning the barcode
178 in the 3' UTR of EGFP (Fig. S3)

179
180 To evaluate the quality of the designed oligonucleotides and the representation of individual
181 barcodes, we sequenced the cloned oligonucleotide library (i.e. prior to reinsertion of the mP +
182 EGFP cassette) to a depth of 19.2 million paired-end consensus sequences, 52.6% of which had
183 the expected length. Analysis of these data showed that most molecular copies of a given
184 oligonucleotide are correct, that synthesis errors are distributed evenly along the designed insert
185 sequence, and that single base deletions dominate the observed errors (Fig. S4A). Nonetheless,
186 there was substantial non-uniformity in the library (Fig. S4B). While 90.5% of the 244,000
187 designed barcodes were observed at least once amongst 11.0 million full-length barcodes
188 sequenced, their abundance is sufficiently dispersed that we estimated that a subset of 56-67% of
189 the designed oligonucleotides would be propagated when maintaining a library complexity of
190 350,000-600,000 clones.

191 Chromosomally integrated versus episomal lentiMPRA

192
193
194 We next sought to directly compare the functional activities of the 2,236 candidate liver enhancer
195 sequences in a chromosomally integrated versus an episomal context. To this end, we packaged
196 the lentiMPRA library with either a wild-type integrase (WT-IN) or a mutant integrase (MT-IN),
197 with the latter allowing for the production of non-integrating lentivirus and transient transgene
198 expression from non-integrated DNA (Leavitt et al. 1996; Nightingale et al. 2006) (Fig. 1A).
199 Because the integrase is not encoded by the lentiMPRA library, this experimental design allows
200 us to test the same exact library in both integrated and non-integrated contexts.

201
202 To optimize conditions and reduce background of unintegrated lentivirus in the integrating
203 lentivirus prep, we utilized our positive control virus (pLS-SV40-mP) that was packaged with WT-
204 IN and MT-IN, and examined the viral titer by qPCR for three different volumes (1, 5 and 25 μ l

205 per well of a 24-well plate) at four different time points (2-5 days post infection). For the lower
206 volumes (1 and 5 ul), we observed a substantial reduction in total virus amounts at day 4 for both
207 MT-IN and WT-IN that stabilized in the WT-IN only (Fig. S5A). This suggests that the non-
208 integrated virus declines at this time point, similar to what was previously reported (Butler et al.
209 2001). For the high volume (25 ul), we did not observe a substantial reduction or stabilization for
210 MT-IN and WT-IN respectively until day 5 (Fig. S5A), suggesting that high amounts of virus
211 would make it difficult to distinguish between integrated and non-integrated virus. We thus
212 decided to obtain DNA/RNA from the cells with the WT-IN liver enhancer library at day 4 when
213 they have an estimated 50 viral particles/cell and the MT-IN library at day 3 when they had an
214 estimated 100 viral particles/cell. The total copy number of viral DNA in the cells infected with
215 the liver enhancer libraries was validated by qPCR (Fig. S5B). During human immunodeficiency
216 virus (HIV) infection, non-integrating virus represents a major portion of the virus at early
217 infection time points and includes linear DNA that is rapidly degraded along with circular DNA
218 containing terminal repeats (1-LTRc and 2-LTRc) (Munir et al. 2013). We further confirmed the
219 copy number of non-integrated virus at our assayed time points by carrying out a qPCR on 2-
220 LTRc, observing the expected low and high amounts of non-integrated virus with WT-IN and MT-
221 IN, respectively (Fig. S5B).

222

223 lentiMPRA on 2,236 candidate liver enhancer sequences

224

225 We recovered RNA and DNA from both WT-IN and MT-IN infections (three replicates each
226 consisting of independent infections with the same library), amplified barcodes, and performed
227 sequencing (Illumina NextSeq). We used both the forward and reverse reads to sequence the 15
228 bp reporter barcodes and obtain consensus sequences. We obtained an average of ~4.1 million raw
229 barcode counts for DNA and an average of ~26 million raw barcode counts for RNA. Across
230 replicates and sample types, 97% of barcodes were the correct length of 15bp. The number of
231 unique sequences was on average ~450,000 for DNA and ~1.2 million for RNA. When clustering
232 sequences with up to one substitution relative to a programmed barcode, the average number of
233 unique sequences reduced to ~280,000 for DNA and ~700,000 for RNA. We speculate that our
234 RNA readouts are impacted by sequence errors to a greater extent due to the reverse transcriptase
235 (RT) step.

236

237 We matched the observed barcodes against the designed barcodes and normalized RNA and DNA
238 for different sequencing depths in each sample by dividing counts by the sum of all observed
239 counts and reporting them as counts per million. Only barcodes observed at least once in both
240 RNA and DNA of the same sample were considered. Subsequently, RNA/DNA ratios were
241 calculated. The average Spearman's rho for DNA counts of the three integrase mutant (MT)
242 experiments was 0.907, and for RNA counts of the MT experiments was 0.982. The average
243 Spearman's rho values for the wild-type integrase (WT) experiments were 0.864 and 0.979 for
244 DNA and RNA, respectively. These correlations were determined for barcodes observed in pairs

245 of replicates. Scatter plots for the MT and WT experiments are shown in Fig. S6 and Fig. S7,
246 respectively.

247
248 While the DNA and RNA counts for individual barcodes are highly correlated between
249 experiments, the noise of each measure results in a poor correlation of RNA/DNA ratios (Fig. S6,
250 Fig. S7). However, there are on average 59-62 barcodes per candidate enhancer sequence (insert)
251 in each replicate (out of 100 barcodes programmed on the array, with ~40% lost during cloning as
252 discussed above) (Fig. S8). To reduce noise, we summed up the RNA or DNA counts across all
253 associated barcodes for each insert observed in a given experiment and recalculated RNA/DNA
254 ratios (Fig. S9). Pairwise-correlations of DNA and RNA counts of replicates are very high (average
255 Spearman's rho MT-RNA 0.996, MT-DNA 0.994, WT-RNA 0.997 and WT-DNA 0.991). Fig. 2
256 shows scatter plots and correlation values for per-insert RNA/DNA ratios for the MT and WT
257 experiments. RNA/DNA ratios show markedly improved reproducibility after summing across
258 barcodes, with an average Spearman's rho of 0.908 (MT) and 0.944 (WT). In all pairwise
259 comparisons of replicates, the integrated (WT) MPRA experiments exhibit a broader dynamic
260 range and greater reproducibility than the episomal (MT) MPRA experiments. We also explored
261 how stable the correlation of RNA/DNA ratios is between replicates by down-sampling the number
262 of barcodes per insert or specifying an exact number of barcodes per insert (Fig. S10). Again, the
263 WT experiments show greater reproducibility, especially for inserts represented by fewer
264 independent barcodes.

265
266 To combine replicates, we normalized the RNA/DNA ratios for inserts observed in each replicate
267 by dividing by their median, and then averaged this normalized RNA/DNA ratio for each insert
268 across replicates (red box in Fig. 2; Fig. 3A). Fig. 3A shows scatter plots of the resulting MT and
269 WT RNA/DNA ratios colored by the type of insert and/or transcription factors considered in the
270 design (Fig. S11 shows RNA/DNA ratio ranges by type of insert). As noted above, we observe a
271 broader dynamic range in the WT experiment. Furthermore, the Spearman correlation between
272 MT and WT is 0.792, which is considerably lower than the correlation observed when correlating
273 replicates of the same experimental type (Spearman correlation of 0.908 (MT) and 0.944 (WT)).
274 This is also the case in pairwise comparisons of MT versus WT replicates (i.e. prior to combining
275 replicates) (yellow boxes in Fig. 2). Overall, these results show that there are substantial
276 differences in regulatory activity between identical sequences assayed in an integrated vs. episomal
277 context.

278
279 Importantly, we can see clear separation of positive and negative controls. Fig. 3B and 3C display
280 RNA/DNA ratios obtained for the highest and lowest SRESs in the MT and WT experiments
281 compared to their previously measured effects in HepG2. While the highest and lowest SRESs are
282 well separated in both experiments (Kolmogorov-Smirnov and Wilcoxon Rank Sum p-values
283 below $2.2E-16$), the WT experiment separates the highest and lowest SRE controls slightly better
284 than the MT experiment (Kolmogorov-Smirnov test D 0.97 vs 0.95, Wilcoxon Rank Sum test W

285 9951 vs 9937). Further, relative to the 90th percentile of SRES negative controls in each
286 experiment, a greater proportion of candidate enhancer sequences are active with integration (36%
287 in WT vs. 28% in MT; Table S1).

288
289 We next sought to assess whether any of our design categories (i.e. types 1-4 defined above,
290 reflecting subsets of candidate enhancers with coincident liver TF and/or chromatin remodeler
291 ChIP-seq peaks) might underlie the observed differences. However, none of these design
292 categories were meaningfully explanatory of enhancer activity or were predictive of differences
293 between MT vs. WT (Figs. 3A and S11).

294 295 ENCODE and other genomic annotations that predict enhancer activity

296
297 Considering that our design categories were predictive of enhancer activity in neither episomal nor
298 chromosomally based MPRA, nor of the differences between them, we explored whether other
299 genomic annotations, some numerical and other categorical (Supplementary File 3), were
300 predictive of our results in HepG2 cells. The performance of individual numerical annotations for
301 predicting the observed activity of candidate enhancer sequences are shown in Fig. 4. We use
302 Kendall's tau, a non-parametric rank correlation that is more conservative than Spearman's rho,
303 because of the large number of zero-values in our annotations which can result in artifacts from
304 ties with Spearman's rho. In contrast with our design bins, many genomic annotations are observed
305 to predict enhancer activity in both the WT and MT experiments. Across the board, annotations
306 correlate better with the WT than the MT results, suggesting that integrated activity read-outs (WT)
307 correlate better with endogenous functional genomic signals (e.g. ChIP-seq data) than do episomal
308 activity read-outs (MT).

309
310 The most highly predictive numerical annotations, in both types of experiments, are HepG2 ChIP-
311 seq datasets of JUND (Transcription Factor Jun-D) and FOSL2 (FOS-Like Antigen 2), consistent
312 with a previous MPRA study which also highlighted the role of these transcription factors in
313 HepG2 cells (Savic et al. 2015). For chromosomally based MPRA (WT), the number of
314 overlapping ENCODE ChIP-seq peaks (TFBS) and the average ENCODE ChIP-seq signal
315 (TFBSPeaks) as measured across different cell-lines also rank amongst the more highly predictive
316 annotations. However, these same features are the most discrepant with MT; that is, substantially
317 less predictive of episomal MPRA. Of note, the highest observed T^2 for an individual annotation
318 is only 0.034 (MT) and 0.058 (WT), leaving a large proportion of the variation in rank order
319 unexplained and highlighting the need for a model combining annotations and other available
320 information (see below).

321
322 We also analyzed how categorical annotations might predict the results of episomal and
323 chromosomal enhancer reporter assays (Fig. S12-S15). Most of these annotations were derived for
324 HepG2 cells by the ENCODE project. However, none of the cell-type specific categorical

325 annotations (ChromHMM (Ernst and Kellis 2012), SegWay (Hoffman et al. 2012) and Open
326 Chromatin annotation) were predictive of the measured RNA/DNA ratios. The multi-cell-type and
327 higher resolution (25 vs 5 level) SegWay chromatin segmentation was most predictive of the
328 measured RNA/DNA ratios. Here, sequences annotated as TSS (transcription start sites) exhibited
329 the highest expression while sequences annotated as D (genomic death zones) exhibited the lowest
330 expression. We note that potential promoters (defined as sites within 1kb of a TSS) comprise ~9%
331 of all non-control sequences (208/2,236) and are enriched in type 3 (49/90) and type 4 (35/87)
332 sequences. We also see the highest proportion of active sequences (where ‘active’ is defined
333 relative to the 90th percentile of SRES negative controls) in the type 3 and 4 categories, even when
334 excluding promoters (Table S1; see also Fig. S11).

335

336 Sequence-based predictors of functional activity

337

338 We next assessed the ability of sequence-based models to predict functional activity of our assayed
339 sequences. Ghandi, Lee *et al.* (Ghandi et al. 2014) introduced a “gapped k-mer” approach for
340 identifying active sequences in a specific cell-type from ENCODE ChIP-seq peaks and matched
341 control sequences (gkm-SVM). The original publication trained models for individual binding
342 factors from up to 5,000 ChIP-seq peaks and the same number of random control sequences. We
343 collected all training data that Ghandi, Lee *et al.* used for HepG2, obtaining ~225,000 unique peak
344 sequences as well as controls, and trained a combined, sequence-based model for predicting ChIP-
345 seq peaks in HepG2 cells (see Methods). Based on a set-aside test dataset, the resulting model had
346 a specificity of 71.8%, a sensitivity of 88.8% and a precision of 75.9% for separating ChIP-seq
347 peak sequences from random control sequences.

348

349 We applied this model to our 171 bp candidate enhancer sequences, and asked how well the
350 resulting gkm-SVM scores correlated with the RNA/DNA ratios obtained for the MT and WT
351 experiments (Fig. 5A-B). The combined gkm-SVM HepG2 model results in a Spearman’s R^2 of
352 0.082 and 0.128, for MT and WT respectively. This is comparable to the best results obtained for
353 individual genomic annotations described before (MT Kendall’s T^2 of 0.038 for gkm-SVM score
354 vs 0.034 for the best individual annotation described before and WT Kendall’s T^2 of 0.060 vs
355 0.058, respectively). However, we note that the correlation with the gkm-SVM model is at least
356 partially driven by the synthetic control sequences, which can be scored with the sequence-based
357 model but not with the genomic annotations. When excluding all control sequences, Spearman’s
358 R^2 values drop from 0.082 to 0.041 and from 0.128 to 0.076 for MT and WT, respectively. As
359 such, there are a few ENCODE-based annotations which outperform the sequence-based gkm-
360 SVM model, namely summaries of JUND/FOSL2 HepG2 ChIP-seq peaks, the number of
361 overlapping ChIP-seq peaks (TFBS) or the average ChIP-seq signal (TFBSPeaks) measured across
362 multiple ENCODE cell-types.

363 Combining annotations and sequence information to predict enhancer activity

364

365 We next sought to combine information across multiple annotations to better predict enhancer
366 activity. We fit Lasso linear models and selected the Lasso tuning parameter value by cross-
367 validation (CV). Scatter plots as well as correlation coefficients were also obtained in a CV setup
368 (see Methods). We built models with all the genomic annotations described above (including the
369 categorical annotations as binary features) as well as with and without the sequence-based gkm-
370 SVM score from scaled and centered annotation matrixes (Fig. S16-18). SRESs and other controls
371 were naturally excluded, as they are largely synthetic sequences and therefore missing genomic
372 annotations. The resulting linear models were considerably more predictive of WT ratios than MT
373 ratios (e.g. CV Spearman R^2 of 0.272 WT vs. 0.146 MT; CV Pearson R^2 of 0.307 WT vs. 0.193
374 MT). Including gapped-kmer SVM scores in the models improved performance further (CV
375 Spearman R^2 of 0.298 WT vs. 0.158 MT; CV Pearson R^2 of 0.330 WT vs. 0.206 MT). We noticed
376 that gapped-kmer SVM scores were assigned the largest model coefficients in both WT and MT
377 models when they were included (Fig. S18). Thus, while reasonably performing models are
378 obtained from genomic annotations, the sequence-based gkm-SVM scores appear to capture
379 independently predictive information.

380

381 We therefore decided to further explore sequence-based models and turned to the faster and low-
382 memory consumption LS-GKM implementation of gkm-SVM (Lee 2016). We trained models
383 from each of the 64 narrow-peak ChIP-seq datasets for which we had included summary statistics
384 for the annotation matrix above (see Methods). We then asked how well the LS-GKM scores
385 generated by each of these 64 models predicted the results of the lentiMPRA experiments.
386 Although the scores now correspond to sequence-based models of ChIP-seq peaks rather than the
387 ChIP-seq peaks themselves, we once again observed the highest Spearman R^2 values for the
388 individual factors JUND (0.117 WT/0.055 MT) and FOSL2 (0.105 WT/0.053 MT), and these are
389 also the factors that show the largest differences in predictive value for WT vs. MT (Fig. S19). As
390 such, sequence-based models of binding by these two factors as well as other individual factors
391 exceed the performance of the pooled gkm-SVM sequence model.

392

393 We fit Lasso linear models from the TF-specific LS-GKM SVM scores in order to predict the
394 measured activities. The combined MT model (using 35 individual scores) achieves a CV
395 Spearman R^2 of 0.134 (CV Pearson R^2 of 0.169), and the combined WT model (using 39 individual
396 scores) of 0.231 (CV Pearson R^2 of 0.263) (Fig. S20). This still falls short of models obtained
397 purely from genomic annotations as described above. To test whether multiple ChIP-seq datasets
398 should be combined in a sequence model rather than combining individual model scores in a linear
399 model to improve prediction, we also trained LS-GKM models based on the peak sequences of the
400 35 (MT) and 39 (WT) scores selected by Lasso models as well as the top 5 and top 10 coefficients
401 in the Lasso models for MT and WT. However, model performance only increased for combining

402 small numbers of peak sets while combining all peaks in one sequence model reduces overall
403 performance (Table S2).

404
405 Finally, when we used both genomic annotations and the individual LS-GKM scores in a single
406 linear model to predict the measured activities, performance increased to a CV Spearman R^2 of
407 0.171 (MT; Pearson R^2 of 0.212) and 0.314 (WT; CV Pearson R^2 of 0.347). These are our highest
408 performing models predicting the activities of candidate enhancer sequences for both the
409 episomally and chromosomally encoded MPRA experiments (Fig. 5C-D).

410 411 **Discussion**

412
413 In this work, we report the first systematic comparison of episomal and chromosomally integrated
414 reporter assays. Key aspects of our approach include: (1) Lentivirus-based MPRA or lentiMPRA,
415 which can be used to in an episomal or integrated context by toggling whether a mutant vs. wild-
416 type integrase is used, and can furthermore be used in a wide variety of cell types, including
417 neurons; (2) The use of numerous barcodes per candidate enhancer sequence, which results in
418 highly reproducible measurements of transcriptional activation; and (3) Extensive predictive
419 modeling of our results, with the implicit assumption that a reasonable measure of a reporter
420 assay's biological relevance is the extent to which it is correlated with endogenous genomic
421 annotations.

422
423 We find that the results of integrated reporter assays are more reproducible, robust and biologically
424 relevant than episomal reporter assays. These conclusions are supported by the following
425 observations: (1) We observed consistently greater reproducibility and dynamic range for the WT
426 replicates as compared with the MT replicates. (2) The correlation of WT vs. MT replicates
427 (Spearman correlation of 0.792) was substantially lower than for WT vs. WT (0.944) or MT vs.
428 MT (0.908), with clear systematic differences between the integrated and episomal contexts that
429 exceed technical noise (Fig. 2, Fig. S9). (3) The WT experiments were consistently more correlated
430 with and more predictable by genomic annotations, which are based on biochemical marks
431 measured in these sequences' native genomic contexts. (4) Many genomic annotations
432 significantly predict the results of the WT but not the MT experiments.

433
434 Of note, we observed generally higher levels of expression with integrated reporters (Fig. 3A and
435 Table S1), consistent with previous findings that showed higher reporter gene levels for integrating
436 relative to non-integrating HIV-1 (Gelderblom et al. 2008; Thierry et al. 2016). However, it is
437 worth noting that we used a lentivirus (not HIV-1) with a self-inactivating (SIN) LTR, which lacks
438 viral promoters or enhancers, potentially influencing these expression differences. Results from
439 hydrodynamic tail vein assays (which delivers reporter constructs into the mouse liver) also show
440 that when chromatinized plasmid DNA leads to higher expression levels than naked plasmid DNA
441 (Kamiya et al. 2013). For HIV-1, both integrating and non-integrating HIV-1 viral DNA are

442 associated with histones (Kantor et al. 2009), which is probably also the case for our lentiviral
443 vector. However, even if the lentiviral episome is chromatinized, there remain myriad potential
444 causes for the observed differences in expression, including differences in H1 stoichiometry,
445 nucleosome positioning, cooperative TF binding (Hebbar and Archer 2007; Hebbar and Archer
446 2008), and/or nuclear location (Jeong and Stein 1994).

447
448 The number of overlapping ENCODE ChIP-seq peaks was one of the most strongly predictive
449 annotations for our integrated sequences (Fig. 4, left). Interestingly, in experiments previously
450 performed on the *MMTV* promoter it was observed that non-integrating constructs could not
451 adequately assess cooperative TF binding due to differences in H1 stoichiometry and nucleosome
452 positioning (Hebbar and Archer 2007), which may relate to the fact that this multiple TF binding
453 was also one of the most differentiating annotations between the WT vs. MT experiments (Fig. 4,
454 right). Specific TF ChIP-seq-based sequence models that are similarly differentiating (Fig. S19)
455 include JUND, FOSL2, ATF3 and ELF1, which are known to interact and form complexes. Jun
456 and Fos family members form the heterodimeric protein complex AP-1, which regulates gene
457 expression in response to various stimuli including stress (Hess et al. 2004) and in the liver has
458 known roles in hepatogenesis (Hilberg et al. 1993) and hepatocyte proliferation (Alcorn et al.
459 1990). AP-1 is known to form complexes with several additional protein partners (Hess et al.
460 2004), including ATF proteins such as ATF3 (Hai and Curran 1991) and ELF1, an ETS
461 transcription factor (Bassuk and Leiden 1995). Of note, while lentivirus infection can induce stress
462 potentially leading to increased expression of AP-1 and related factors, these same TFs were less
463 predictive in MT-infected sequences, and furthermore the ChIP-seq datasets were generated on
464 cells in normal physiological conditions. Combined, these findings suggest that differences in
465 cooperative TF binding, possibly involving TFs including JUND, FOSL2, ELF1 and ATF3, might
466 drive differences in the results of integrated vs. episomal reporter assays.

467
468 Using single-feature models, we also systematically evaluated more than 400 genomic annotations
469 and sequence models to explore which are significantly predictive of expression in the integrated
470 and/or episomal lentiMPRA experiments (Figure S21). Consistent with our other analyses, there
471 are many more annotations that are significantly predictive of the integrated (WT) assay but not
472 the episomal (MT) assay. These include several annotations related to histone acetylation
473 (HDAC2, EP300, and ZBTB7A) as well as a factor with increased liver expression and which is
474 associated with adipogenesis, gluconeogenic and hematopoiesis (CEBPB) (Tsukada et al. 2011).
475 Interestingly, there are also a number of annotations which are only significantly predictive of the
476 episomal assay, including BRCA1 ChIP-seq peaks and SIN3 transcriptional regulator family
477 member B (SIN3B) binding motifs.

478
479 A contemporary challenge for our field is how to best identify, prioritize, and functionally validate
480 *cis*-regulatory elements, especially enhancers. To address this, we envision a virtuous cycle, in
481 which annotation and/or sequence-based models are used to nominate candidate enhancer

482 sequences for validation, these candidates are tested in massively parallel reporter assays, and then
483 the results are used to improve the models which in turn results in higher quality nominations.
484 Eventually, this will lead to not only a catalog of validated enhancers but also a deeper mechanistic
485 understanding of the relationship between primary sequence, transcription factor binding, and
486 quantitative enhancer activity. In this study, our best performing model achieves a Pearson's R^2 of
487 0.347 in predicting the results of the integrated lentiMPRA, with both genomic annotations and
488 sequence-based models providing independent information. Of note, these are quantitative
489 predictions of activity, a more challenging task than simply categorizing enhancers vs. non-
490 enhancers. Although far from perfect, we are able to garner insights into the determinants of
491 enhancer function, and may be able to use this model to select a much larger number of candidate
492 enhancer sequences for testing and further modeling.

493
494 As our field scales MPRA to characterize very large numbers of candidate enhancers, it is
495 obviously critical that the reporter assays are as reflective as possible of endogenous biology. Our
496 results directly test a longstanding concern about episomal reporter assays, and suggest that there
497 are substantial differences between the integrated and episomal contexts. Furthermore, based on
498 the fact that their output is more correlated with genomic annotations, we infer that integrated
499 reporter assays are more reflective of endogenous enhancer activity. This fits with our expectation,
500 as both the integrated reporter and endogenous enhancers reside within chromosomes as opposed
501 to episomes. We urge caution in the interpretation of the results of all reporter assays, and that
502 integrated reporter assays such as lentiMPRA be used where possible.

503

504 **Methods**

505

506 Lentivirus enhancer construct generation

507

508 To generate the lentivirus vector (pLS-mP), a minimal promoter sequence, which originates from
509 pGL4.23 (Promega), including an *SbfI* site was obtained by annealing of oligonucleotides (Sense:
510 5'- CTAGACCCTGCAGGCACTAGAGGGTATATAATGGAAGCTCGACTTCCAGCTTGG
511 CAATCCGGTACTGTA-3', Antisense: 5'- CCGGTACAGTACCGGATTGCCAAGCTGGAA
512 GTCGAGCTTCCATTATATACCCTCTAGTGCCCTGCAGGT-3'; *SbfI* site is underlined), and
513 subcloned into *XbaI* and *AgeI* sites in the pLB vector (Addgene 11619; (Kissler et al. 2006))
514 replacing the U6 promoter and CMV enhancer/promoter sequence in the vector. To generate pLS-
515 mP-SV40, the SV40 enhancer sequence was amplified from pGL4.13 (Promega) using primers
516 (Forward: 5'- CAGGGCCCGCTCTAGAGCGCAGCACCATGGCCTGAA-3', Reverse: 5'-
517 TGCCTGCAGGTCTAGACAGCCATGGGGCGGAGAATG-3') and inserted into *XbaI* site in
518 the vector using In-Fusion (Clontech). pos1, pos2, neg1, and neg2 sequences were amplified from
519 human (pos1, neg2, pos2) or mouse (neg1) genome, and inserted into *EcoRV* and *HindIII* site in
520 pGL4.23 (Promega). Primers used are shown in Table S3 and the annotated plasmid sequence file
521 is available as Supplementary File 1.

522

523 Library sequence design

524

525 We picked 171bp candidate enhancer sequences based on ChIP-seq peaks calls for HepG2. We
526 used narrow peak calls for DNA binding proteins/transcription factors (FOXA1, FOXA2, HNF4A,
527 RAD21, CHD2, SMC3 and EP300) and wide peak calls for histone marks (H3K27ac). We
528 downloaded the call sets from the ENCODE portal (Sloan et al. 2016)
529 (<https://www.encodeproject.org/>) with the following identifiers: ENCFF001SWK,
530 ENCFF002CKI, ENCFF002CKJ, ENCFF002CKK, ENCFF002CKN, ENCFF002CKY,
531 ENCFF002CUS, ENCFF002CTX, ENCFF002CUU, ENCFF002CKV, and ENCFF002CUN. We
532 defined four classes of sites: 1) Regions centered over peak calls of ≤ 171 bp for FOXA1, FOXA2
533 or HNF4A that overlap H3K27ac and EP300 calls as well as at least one of three chromatin
534 remodeling factors RAD21, CHD2 or SMC3. 2) Regions like in 1, but with no remodeling factor
535 overlap. 3) Regions of 171 bp centered in an EP300 peak overlapping H3K27ac as well as at least
536 one of three chromatin remodeling factors RAD21, CHD2 or SMC3, but without peaks in FOXA1,
537 FOXA2 or HNF4A. 4) Regions like in 3 but with no remodeling factor overlap. Sites of type 1 and
538 2 involving HNF4 were the most abundant sites and we used those to fill-up our design after
539 exhausting other target sequences.

540

541 Potential 171bp target sequences were inserted into a 230bp-oligo backbone with a 5'-flanking
542 sequence (15bp, AGGACCGGATCAACT), 14bp-spacer sequence (CCTGCAGGGAATTC),
543 15bp designed tag sequences (see below) and a 3'-flanking sequence (15bp,
544 CATTGCGTGAACCGA) (Supplementary Fig. Y). Sequences were checked for *SbfI* and *EcoRI*
545 restriction sites after joining the target sequence with the 5'-flanking sequence and the spacer
546 sequence. Such potential target sequences were discarded.

547

548 In our final array design, we included 2,440 different target sequences each with 100 different
549 barcodes (i.e. a total of 244,000 oligos). These included the highest 100 and lowest 100 synthetic
550 regulatory element (SRE) sequences identified by Smith RP et al. (Smith et al. 2013), 4 control
551 sequences (neg1 MGSCv37 chr19 35,531,983-35,532,154, neg2 GRCh37 chr5 172,177,151-
552 172,177,323, pos1 GRCh37 chr3 197,439,136-197,439,306, pos2 GRCh37 chr19 35,531,984-
553 35,532,154) which we tested using Luciferase assays in the HepG2 cell line (Fig. S2), 1,029 type
554 1 inserts (202 FOXA1, 180 FOXA2, 464 HNF4A, 120 FOXA1&FOXA2, 33 FOXA1&HNF4A,
555 17 FOXA2&HNF4A, 13 FOXA1&FOXA2&HNF4A), 1,030 type 2 inserts (195 FOXA1, 174
556 FOXA2, 470 HNF4A, 126 FOXA1&FOXA2, 31 FOXA1&HNF4A, 20 FOXA2&HNF4A, 14
557 FOXA1&FOXA2&HNF4A), 90 type 3 inserts and 87 type 4 inserts.

558

559 Tag sequences of 15bp length were designed to have at least two substitutions and one 1bp-
560 insertion distance to each other. Homopolymers of length 3 bp and longer were excluded in the
561 design of these sequences, and so were ACA/CAC and GTG/TGT trinucleotides (bases excited
562 with the same laser during Illumina sequencing). More than 556,000 such barcodes were designed

563 using a greedy approach. The barcodes were then checked for the creation of *SbfI* and *EcoRI*
564 restriction sites when adding the spacer and 3'-flanking sequences. From the remaining sequences,
565 a random subset of 244,000 barcodes was chosen for the design. The final designed oligo
566 sequences are available in Supplementary File 2.

567

568 Generation of MPRA libraries

569

570 The lentiviral vector pLS-mP was cut with *SbfI* and *EcoRI* to temporarily liberate the minimal
571 promoter and EGFP reporter gene. Array-synthesized 230bp oligos (Agilent Technologies)
572 containing an enhancer, spacer, and barcode (Fig. S3) were amplified with adaptor primers
573 (pLSmP-AG-f and pLSmP-AG-r) that have overhangs complementary to the cut vector backbone
574 (Table S3), and the products were cloned using NEBuilder HiFi DNA Assembly mix (E2621). The
575 adaptors were chosen to disrupt the original *SbfI* and *EcoRI* sites in the vector. The cloning reaction
576 was transformed into electrocompetent cells (NEB C3020). Multiple transformations were pooled
577 and midi-prepped (Chargeswitch Pro Filter Plasmid Midi Kit, Invitrogen CS31104). This library
578 of cloned enhancers and barcodes was then cut using *SbfI* and *EcoRI* sites contained within the
579 spacer, and the minimal promoter and *EGFP* that were removed earlier were reintroduced via a
580 sticky end ligation (T4 DNA Ligase, NEB M0202) between the enhancer and its barcode. These
581 finished vectors were transformed and midi-prepped as previously mentioned.

582

583 Quality control of designed array oligos

584

585 Before inserting the minimal promoter and EGFP reporter gene, the plasmid library was sampled
586 by high-throughput sequencing on an Illumina MiSeq (206/200+6 cycles) to check for the quality
587 of the designed oligos and the representation of individual barcodes (sequencing primers are
588 pLSmP-AG-seqR1, pLSmP-AG-seqIndx, and pLSmP-AG-seqR2; Table S3). We sequenced the
589 target, spacer and tag sequences from both read ends and called a consensus sequence from the
590 two reads. We obtained 19.2 million paired-end consensus sequences from this sequencing
591 experiment. 52.6% of those sequences had the expected length, 26.1% of sequences were 1bp short
592 and 8.9% were 2bp short (summing up to 87.6%). Only 1.6% of sequences showed an insertion of
593 1bp. These results are in line with expected dominance of small deletion errors in oligo synthesis.
594 We aligned all consensus sequences back to all designed sequences using BWA MEM (Li and
595 Durbin 2009) with parameters penalizing soft-clipping of alignment ends (-L 80). We consensus
596 called reads aligning with the same outer alignment coordinates and SAM-format CIGAR string
597 to reduce the effects of sequencing errors. We analyzed all those consensus sequences based on at
598 least three sequences with a mapping quality above 0. We note that substitutions are removed in
599 the consensus calling process if the correct sequence is the most abundant sequence. Among these
600 992,513 consensus sequences, we observe instances of 91% designed oligos and 78% of oligos
601 with one instance matching the designed oligo perfectly. Across all consensus sequences the
602 proportion of perfect oligos is only 19%, however the proportion vastly increases with the number

603 of observations (69% at 20 counts, 99% at 40 counts; Table S4). These observations are in
604 agreement with most molecular copies of an oligo being correct, in combination with high
605 representation differences in the library. Supplementary Fig. S4A shows the distribution of
606 alignment differences (as a proxy for synthesis errors) along the designed oligo sequences. Errors
607 are distributed evenly along the designed insert sequence, with deletions dominating the observed
608 differences. We observe that at some positions the deletion rate is reduced while the insertion rate
609 is increased. We speculate that this might be due to certain sequence contexts.

610

611 Limited coverage of designed oligos in MPRA libraries

612

613 From the analysis of oligo quality and oligo abundance above, we saw a first indication of the
614 existence of a wide range of oligo abundance and that more frequent sequences tend to match the
615 designed sequences perfectly (see Table S4). We characterized the abundance of oligos further and
616 looked at the consequences that this has for generating libraries of lentivirus constructs with
617 limited complexity (due to the transformation of a limited number of bacteria). Rather than looking
618 at full length oligos, we focused only on the tag sequences. Tag sequences were identified from
619 the respective alignment positions of the alignments created above. To match the RNA/DNA count
620 data analysis (see below), we only considered barcodes of 15bp length (10.96M/57.0%, similar to
621 the proportion of correct length sequences above). Of those 10.96M barcodes, 345,247 different
622 sequences are observed. We clustered (dnaclust (Ghodsi et al. 2011)) the remaining sequences
623 allowing for one substitution and selecting the designed or most abundant sequence (reducing to
624 238,206 different sequences). The clustered sequences were matched against the designed
625 barcodes (217,176 sequences, 99.2% of counts). The distribution of the abundance of these
626 barcodes is available in Supplementary Fig. S4B. We used those counts to simulate sampling from
627 this over dispersed pool of sequences, as done when taking a sample of plasmids infusing the
628 reporter gene and minimal promoter and again transforming the resulting plasmids. We sampled
629 10 times and averaged the number of unique designed barcodes: 150k clones – 87,944 unique
630 barcodes, 250k clones – 116,297 unique barcodes, 350k clones – 135,222 unique barcodes, 500k
631 clones – 154,090 unique barcodes, 600k clones – 163,831 unique barcodes, 750k clones – 172,770
632 unique barcodes and 1M clones – 183,685 unique barcodes. Thus, even for high sampling depth
633 only a subset of barcodes will be captured in the final library. We observe on average 145,876
634 different barcodes which is concordant with more than 430k clones going into the lenti
635 construction.

636

637 Cell culture and GFP / luciferase assays

638

639 HepG2 cells were cultured as previously described (Smith et al. 2013). K562, H1-ESC, HeLa-S3,
640 T-47D and Sk-n-sh cells were culture as previously described (Dunham I et al. 2012). Sk-n-sh
641 cells were treated with 24 uM *all trans*-retinoic acid (Sigma) to induce neuronal differentiation.
642 K562, H1-ESC, HeLa-S3, T-47D and Sk-n-sh were treated with retinoic acid and infected with

643 pLS-mP or pLS-SV40-mP lentivirus along with 8 $\mu\text{g}/\text{ml}$ polybrene, and incubated for 2 days, when
644 they have an estimated 30, 60, 90, 90, and 90 viral particles/cell, respectively. The number of viral
645 particles/cell was measured as described below. For the four control sequences (two negatives and
646 two positives) luciferase assay, we amplified the controls from the designed oligo pool (primer
647 sequences available in Table S5) and inserted those into the pGL4.23 (Promega) reporter plasmid.
648 2×10^4 HepG2 cells/well were seeded in a 96-well plate. 24 hour later, the cells were transfected
649 with 90 ng of reporter plasmids (pGL4.23-neg1, pGL4.23-neg2, pGL4.23-pos1, and pGL4.23-
650 pos2) and 10 ng of pGL4.74 (Promega), which constitutively expresses Renilla luciferase, using
651 X-tremeGENE HP (Roche) according to the manufacturer's protocol. The X-tremeGENE:DNA
652 ratio was 2:1. Three independent replicate cultures were transfected. Firefly and Renilla luciferase
653 activities were measured as previously described (Smith et al. 2013).

654

655 Lentivirus packaging, titration and infection

656

657 Twelve million HEK293T cells were plated in 15 cm dish and cultured for 24 hours. The cells
658 were co-transfected with 8 μg of the liver enhancer library and 4 μg of packaging vectors using
659 jetPRIME (Polyplus-transfections). psPAX2 that encodes wild-type *pol* was used to generate
660 integrating lentivirus, while pLV-HELP (InvivoGen) that encodes a mutant *pol* was used to
661 generate non-integrating lentivirus. pMD2.G was used for both. The transfected cells were cultured
662 for 3 days and lentivirus were harvested and concentrated as previously described (Wang and
663 McManus 2009).

664

665 To measure DNA titer for the integrating and non-integrating lentivirus library, HepG2 cells were
666 plated at 2×10^5 cells/well in 12-well plates and incubated for 24 hours. Serial volume (0, 1, 5, 25
667 μL) of the lentivirus was added with 8 $\mu\text{g}/\text{ml}$ polybrene, to increase infection efficiency. The
668 infected cells were cultured for 2-5 days and then washed with PBS three times. Genomic DNA
669 was extracted using the Wizard SV genomic DNA purification kit (Promega). Copy number of
670 viral particle per cell was measured as relative amount of viral DNA (WPRE region) over that of
671 genomic DNA (intronic region of *LIPC* gene) by qPCR using SsoFast EvaGreen Supermix
672 (BioRad), according to manufacturer's protocol. PCR primer sequences are shown in Table S3.
673 For the lentiMPRA, 2.4 million HepG2 cells were plated on a 10 cm dish and cultured for 24 hours.
674 The cells were infected with integrating or non-integrating lentivirus libraries along with 8 $\mu\text{g}/\text{ml}$
675 polybrene, and incubated for 4 and 3 days, when they have an estimated 50 and 100 viral
676 particles/cell, respectively. Three independent replicate cultures were infected. The cells were
677 washed with PBS three times, and genomic DNA and total RNA was extracted using AllPrep
678 DNA/RNA mini kit (Qiagen). Copy number of viral particle per cell was confirmed by qPCR and
679 shown in Supplementary Fig. S5B. Messenger RNA (mRNA) was purified from the total RNA
680 using Oligotex mRNA mini kit (Qiagen) and treated with Turbo DNase to remove contaminating
681 DNA.

682

683 RT-PCR, amplification and sequencing of RNA/DNA

684
685 For each replicate, 3x500ng was reverse transcribed with SuperscriptII (Invitrogen 18064-014)
686 using a primer downstream of the barcode (pLSmP-ass-R-i#, Table S3), which contained a sample
687 index and a P7 Illumina adaptor sequence. The resulting cDNA was pooled and split into 24
688 reactions, amplified with Kapa Robust polymerase for 3 cycles using this same reverse primer
689 paired with a forward primer complementary to the 3' end of EGFP with a P5 adaptor sequence
690 (BARCODE_lentiF_v4.1, Table S3). The implemented two-round PCR set-up is supposed to
691 reduce PCR jack-potting effects and allows for incorporating unique molecular identifiers (UMIs),
692 which could be used to correct for other PCR biases in future experiments. PCR products are then
693 cleaned up with AMPure XP beads (Beckman Coulter) to remove the primers and concentrate the
694 products. These products underwent a second round of amplification in 8 reactions per replicate
695 for 15 cycles, with a reverse primer containing only P7. All reactions were pooled at this point,
696 run on an agarose gel for size-selection, and submitted for sequencing. For the DNA, 16x500ng of
697 each replicate was amplified for 3 cycles just as the RNA. First round products were cleaned up
698 with AMPure XP beads, and amplified for another 16 reactions, each for 20 cycles. Reactions were
699 pooled, gel-purified, and sequenced. Sequencing primers are BARCODE-SEQ-R1-V4, pLSmP-
700 AG-seqIndx, and BARCODE-SEQ-R2-V4 for both RNA and DNA barcodes (Table S3).

701
702 RNA and DNA for each of three replicates was sequenced on an Illumina NextSeq instrument
703 (2x26 + 10bp index). The forward and reverse reads on this run each sequenced the designed 15bp
704 barcodes as well adjacent sequence to correctly trim and consensus call barcodes. We obtained a
705 minimum of 2.9M and a maximum of 5.9M raw counts for DNA (average 4.1M) and a minimum
706 of 20.0M and a maximum of 32.3M raw counts for RNA (average 25.6M). Across replicates and
707 sample type, 97% of barcodes were of the correct length of 15bp.

708
709 The number of unique sequences was on average 446k for DNA and 1.2M for RNA. When
710 clustering sequences with one substitution (dnaclust; (Ghodsi et al. 2011)), the average number of
711 unique sequences reduced to 280k for DNA and 697k for RNA. When overlapping the observed
712 with the designed sequences, clustering keeps more counts but reduces the total number of
713 observed barcodes (93.1% vs. 90.3%, 145k vs 151k). We believe this is due to too many errors in
714 barcodes which are sufficiently similar to cause clusters to merge across different designed tag
715 sequences. We therefore dismissed the clustered data and only matched against the designed
716 barcodes. This is further supported by counts being more highly correlated between replicates
717 when using the non-clustered data (Spearman's rho without clustering: DNA replicates 88.6%,
718 RNA replicates 98.0%; with clustering: DNA replicates 85.0%, RNA replicates 94.3%).

719 Replicates, normalization and RNA/DNA ratios

721

722 To normalize RNA and DNA for different sequencing depths in each sample, we divided reads by
723 the sum of observed counts and reported them as counts per million. Only barcodes observed in
724 RNA and DNA of the same sample were considered. Subsequently, RNA/DNA ratios were
725 calculated. We observe that the dynamic range observed in the WT experiments is larger and that
726 the average Spearman's rho is also higher for the WT experiments (44.3% vs. 39.0%). To
727 determine the RNA/DNA ratios per insert, we summed up the counts of all barcodes contributing
728 and determined the ratio of the average normalized counts. We explored how stable the correlation
729 of RNA/DNA ratios is between replicates when limiting the number of barcodes per insert (Fig
730 S10). We limited the maximum number of barcodes considered by (1) randomly down sampling
731 and (2) requiring an exact number of barcodes per insert (i.e. down sampling those with more and
732 excluding those inserts with less barcodes).

733
734 Even though normalized individually, the three replicates of each experiment do not seem to be on
735 the exact same scale (Figure 2 & S9). We therefore chose to divide the RNA/DNA ratios by the
736 median across the technical replicate value before averaging them.

737 738 Predictors of sequence effects

739 To correlate available annotations with the observed sequence activity in HepG2 cells, we
740 downloaded additional narrow peak calls for DNA binding proteins/transcription factors in HepG2
741 from ENCODE data. We obtained call-sets for the following 64 factors: ARID3A, ATF3,
742 BHLHE40, BRCA1, CBX1, CEBPB, CEBPD, CHD2, CTCF, ELF1, EP300, EZH2, FOSL2,
743 FOXA1, FOXA2, FOXK2, GABPA, GATA4, HCFC1, HDAC2, HNF4A, HNF4G, IRF3, JUN,
744 JUND, MAFF, MAFK, MAX, MAZ, MBD4, MXI1, MYBL2, MYC, NFIC, NR2C2, NRF1,
745 POLR2A, POLR2AphosphoS2, POLR2AphosphoS5, RAD21, RCOR1, REST, RFX5, RXRA,
746 SIN3A, SIN3B, SMC3, SP1, SP2, SRF, TAF1, TBP, TCF12, TCF7L2, TEAD4, TFAP4, USF1,
747 USF2, YY1, ZBTB33, ZBTB7A, ZHX2, ZKSCAN1, and ZNF274. Additionally, we downloaded
748 ChromHMM segmentations for HepG2, Open Chromatin State, SegWay, and DHS call-sets from
749 the ENCODE portal (Sloan et al. 2016). From the NIH Roadmap Epigenomics Consortium we
750 obtained RNAseq, DNA methylation, DNase, CAGE, H2A.Z, H3K4me1, H3K4me2, H3K4me3,
751 H3K9ac, H3K9me3, H3K27ac, H3K27me3, H3K36me3, H3K79me2, and H4K20me1 tracks
752 (Kundaje et al. 2015). We also downloaded the Fantom5 Robust Enhancer annotations (Andersson
753 et al. 2014), Fantom5 CAGE data for HepG2 (Forrest et al. 2014), GenoSTAN Enhancer and
754 promoter predictions (<http://i12g-gagneurweb.in.tum.de/public/paper/GenoSTAN/>),
755 enhancerFinder predictions (Erwin et al. 2014), as well as motif scan results and annotated
756 regulatory features from the Ensembl Regulatory Build (Zerbino et al. 2015). For further genome-
757 wide and organismal metrics, we turned to the CADD v1.3 annotation file (Kircher et al. 2014)
758 and extracted local GC and CpG content, SegWay, chromHMM state across the NIH RoadMap
759 cell types, priPhCons, mamPhCons, verPhCons, priPhyloP, mamPhyloP, verPhyloP, GerpN,
760 GerpS, GerpRS, bStatistic, tOverlapMotifs, motifECount, motifEHIPos, TFBS, TFBSPeaks,
761 TFBSPeaksMax, distance to TSS, and the actual CADD score column. We included the number

762 of bases covered by peak calls as well as the average and maximum values across the designed
763 sequences for those metrics. Supplementary File 3 outlines all annotations used.

764

765 Gapped-kmer SVM (gkm-SVM) model of HepG2 activity

766

767 We collected training data of individual ChIP-seq binding factors described by M Ghandi & D Lee
768 et al. (Ghandi et al. 2014) for HepG2 (5k specific ChIP-seq peak regions and the same number of
769 random controls, <http://www.beerlab.org/gkmsvm/>) and removed duplicate sequences, obtaining
770 225k peak sequences as well as matched random controls. Attempting to train a classification
771 model with the gkm-SVM software based on all peak sequences exceeded reasonable memory
772 requirements (>1TB). Therefore, we iteratively reduced the number of training examples and
773 ended up sampling each 50k peak and 50k control sequences for a combined HepG2 sequence
774 model. Based on a test data set (2k sampled from the unused training data set), the obtained model
775 has a specificity of 71.8%, a sensitivity of 88.8% and precision of 75.9% for separating ChIP-seq
776 peak from the random control sequences.

777

778 Linear models integrating individual annotations

779

780 We used the R glmnet package to fit Lasso-penalized linear models to predict RNA/DNA ratios.
781 We used 10-fold cross-validation (cv.glmnet) to determine the Lasso tuning parameter lambda
782 resulting in the minimum squared error. The Lasso forces small coefficients to zero, and thereby
783 performs regression and feature selection simultaneously. Categorical features with K levels were
784 included as K-1 binary columns. We excluded ZNF274 and EZH2 annotations from the model as
785 none of the inserts overlapped with these ChIP-seq tracks. Otherwise missing annotation values
786 were mostly in count features (70.1%) or absence of the conserved block annotation “GerpRS”
787 (27.5%) and thus all these values were imputed to zero. All annotation features were scaled and
788 centered. To report unbiased correlation values and scatter plots between the true and predicted
789 RNA/DNA ratios, we randomly split up our data into 10 folds, trained models using 9 folds and
790 the above identified tuning parameter and then extracted the fitted values after applying the model
791 to the remaining fold.

792

793 Sequence-based LS-GKM models

794

795 LS-GKM (Lee 2016) is a faster and lower memory profile version of gkm-SVM. Its default settings
796 are different from gkm-SVM (e.g. using 11 bases with 7 informative positions rather than 10 bases
797 with 6 informative positions). We applied LS-GKM using parameters corresponding to gkm-SVM
798 (-l 10 -k 6 -d 3 -t 2 -T 4 -e 0.01) as well as default parameters (-T 4 -e 0.01) on the HepG2 training
799 data described for gkm-SVM above (225,327 positive/negative sequences each, 10,000 kept set
800 aside for validation). We also compared performance for using the negative sequences as described
801 for gkm-SVM (Ghandi et al. 2014) versus obtaining negative sequences by permutation of the real

802 sequences maintaining dinucleotide content (Jiang et al. 2008). We found that best results were
803 obtained for LS-GKM defaults in combination with selected negative sequences rather than
804 permuted sequences (Table S6). However, permuted sequences as negative set produced a higher
805 true positive rate and substantially simplify computation. We therefore used permuted sequence
806 sets and ran LS-GKM with default parameters for all models. We extracted genomic sequences
807 (GRCh37) below the 64 ChIPseq peak sets by concatenating multiple call-sets for the same factor
808 and merging overlapping peak regions using bedtools (Quinlan and Hall 2010). We extracted up
809 to 1kb of sequence for each peak, or centered 1kb fragments on the peak for larger peak calls. We
810 chose the model convergence parameter ϵ based on the number of positive training sequences
811 (mean 16600, min. 186, max. 63948) multiplied with 1E-07; investing more training iterations for
812 smaller training data sets.

813
814 We then used Lasso regression (as described above) to create combined models and we also trained
815 LS-GKM models from pooled peak data sets (Table S2). For this purpose, we pooled sequences
816 using the peak data sets underlying the top 5, top 10 and all sequence models selected using Lasso
817 regression.

818

819 Individual feature models

820

821 To explore whether certain annotations are more strongly predictive for either the non-integrated
822 (MT) or integrated (WT) expression measurements (despite the correlations among the
823 annotations), we used the R glm (Generalized Linear Models) implementation to fit 430 linear
824 single coefficient plus intercept models for predicting log₂ RNA/DNA ratios for MT and WT
825 experiments. We report the two-sided p-value for the t-statistic corresponding to the coefficient in
826 the linear model, and used a significance criterion of 0.05 after Bonferroni correction (Figure S21,
827 Table S7).

828

829 **Data access**

830

831 Raw sequencing data, designed oligo sequences and processed count and RNA/DNA ratio data
832 including annotations was submitted to the NCBI Gene Expression Omnibus (GEO) and was
833 assigned accession GSE83894.

834

835

836 **Acknowledgements**

837

838 We thank members of the Ahituv, McManus and Shendure laboratories for helpful discussions
839 and suggestions. Our work was supported by the National Human Genome Research Institute
840 (NHGRI) grant number 1R01HG006768 (NA & JS), NHGRI and National Cancer Institute grant
841 number 1R01CA197139 (GC, DW, NA, JS) and National Institute of Mental Health grant number
842 1R01MH109907 (NA). JS is an investigator of the Howard Hughes Medical Institute.

843

844 **Author's contributions**

845

846 FI, MK, BM, MTM, NA and JS designed experiments; FI and BM performed all wet lab
847 experiments; MK, JS, NA, DMW and GMC outlined data analysis; MK performed data analysis;
848 FI, MK, NA, JS interpreted the experimental results; FI, MK, BM, NA and JS wrote the
849 manuscript.

850

851 References

- 852
- 853 Alcorn JA, Feitelberg SP, Brenner DA. 1990. Transient induction of c-jun during hepatic regeneration.
854 *Hepatology* **11**: 909-915.
- 855 Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, Chen Y, Zhao X, Schmidl C,
856 Suzuki T et al. 2014. An atlas of active enhancers across human cell types and tissues. *Nature* **507**:
857 455-461. doi: 410.1038/nature12787.
- 858 Archer TK, Lefebvre P, Wolford RG, Hager GL. 1992. Transcription factor loading on the MMTV
859 promoter: a bimodal mechanism for promoter activation. *Science* **255**: 1573-1576.
- 860 Arnold CD, Gerlach D, Spies D, Matts JA, Sytnikova YA, Pagani M, Lau NC, Stark A. 2014. Quantitative
861 genome-wide enhancer activity maps for five Drosophila species show functional enhancer
862 conservation and turnover during cis-regulatory evolution. *Nat Genet* **46**: 685-692. doi:
863 610.1038/ng.3009. Epub 2014 Jun 1038.
- 864 Arnold CD, Gerlach D, Stelzer C, Boryn LM, Rath M, Stark A. 2013. Genome-wide quantitative enhancer
865 activity maps identified by STARR-seq. *Science* **339**: 1074-1077. doi: 1010.1126/science.1232542.
866 Epub 1232013 Jan 1232517.
- 867 Banerji J, Rusconi S, Schaffner W. 1981. Expression of a beta-globin gene is enhanced by remote SV40
868 DNA sequences. *Cell* **27**: 299-308.
- 869 Bassuk AG, Leiden JM. 1995. A direct physical association between ETS and AP-1 transcription factors in
870 normal human T cells. *Immunity* **3**: 223-237.
- 871 Butler SL, Hansen MS, Bushman FD. 2001. A quantitative assay for HIV DNA integration in vivo. *Nat*
872 *Med* **7**: 631-634.
- 873 Dunham I Kundaje A Aldred SF Collins PJ Davis CA Doyle F Epstein CB Frietze S Harrow J et al. 2012.
874 An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**: 57-74.
- 875 Ernst J, Kellis M. 2012. ChromHMM: automating chromatin-state discovery and characterization. *Nat*
876 *Methods* **9**: 215-216. doi: 210.1038/nmeth.1906.
- 877 Erwin GD, Oksenberg N, Truty RM, Kostka D, Murphy KK, Ahituv N, Pollard KS, Capra JA. 2014.
878 Integrating diverse datasets improves developmental enhancer prediction. *PLoS Comput Biol* **10**:
879 e1003677. doi: 1003610.1001371/journal.pcbi.1003677. eCollection 1002014 Jun.
- 880 Euskirchen GM, Auerbach RK, Davidov E, Gianoulis TA, Zhong G, Rozowsky J, Bhardwaj N, Gerstein
881 MB, Snyder M. 2011. Diverse roles and interactions of the SWI/SNF chromatin remodeling
882 complex revealed using global approaches. *PLoS Genet* **7**: e1002008. doi:
883 1002010.1001371/journal.pgen.1002008. Epub 1002011 Mar 1002003.
- 884 Faure AJ, Schmidt D, Watt S, Schwalie PC, Wilson MD, Xu H, Ramsay RG, Odom DT, Flicek P. 2012.
885 Cohesin regulates tissue-specific expression by stabilizing highly occupied cis-regulatory modules.
886 *Genome Res* **22**: 2163-2175. doi: 2110.1101/gr.136507.136111. Epub 132012 Jul 136510.
- 887 Forrest AR Kawaji H Rehli M Baillie JK de Hoon MJ Haberle V Lassmann T Kulakovskiy IV Lizio M Itoh
888 M et al. 2014. A promoter-level mammalian expression atlas. *Nature* **507**: 462-470. doi:
889 410.1038/nature13182.
- 890 Gelderblom HC, Vatakis DN, Burke SA, Lawrie SD, Bristol GC, Levy DN. 2008. Viral complementation
891 allows HIV-1 replication without integration. *Retrovirology* **5**:60. doi: 10.1186/1742-4690-1185-1160.
- 892 Ghandi M, Lee D, Mohammad-Noori M, Beer MA. 2014. Enhanced regulatory sequence prediction using
893 gapped k-mer features. *PLoS Comput Biol* **10**: e1003711. doi:
894 1003710.1001371/journal.pcbi.1003711. eCollection 1002014 Jul.
- 895 Ghodsi M, Liu B, Pop M. 2011. DNACLUSt: accurate and efficient clustering of phylogenetic marker
896 genes. *BMC Bioinformatics* **12**:271. doi: 10.1186/1471-2105-1112-1271.
- 897 Hai T, Curran T. 1991. Cross-family dimerization of transcription factors Fos/Jun and ATF/CREB alters
898 DNA binding specificity. *Proc Natl Acad Sci U S A* **88**: 3720-3724.
- 899 Hebbar PB, Archer TK. 2007. Chromatin-dependent cooperativity between site-specific transcription
900 factors in vivo. *J Biol Chem* **282**: 8284-8291.

- 901 Hebbbar PB, Archer TK. 2008. Altered histone H1 stoichiometry and an absence of nucleosome positioning
902 on transfected DNA. *J Biol Chem* **283**: 4595-4601.
- 903 Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, Barrera LO, Van Calcar S, Qu C, Ching
904 KA et al. 2007. Distinct and predictive chromatin signatures of transcriptional promoters and
905 enhancers in the human genome. *Nat Genet* **39**: 311-318.
- 906 Hess J, Angel P, Schorpp-Kistner M. 2004. AP-1 subunits: quarrel and harmony among siblings. *J Cell Sci*
907 **117**: 5965-5973.
- 908 Hilberg F, Aguzzi A, Howells N, Wagner EF. 1993. c-jun is essential for normal mouse development and
909 hepatogenesis. *Nature* **365**: 179-181.
- 910 Hoffman MM, Buske OJ, Wang J, Weng Z, Bilmes JA, Noble WS. 2012. Unsupervised pattern discovery
911 in human chromatin structure through genomic segmentation. *Nat Methods* **9**: 473-476. doi:
912 410.1038/nmeth.1937.
- 913 Jeong S, Stein A. 1994. Micrococcal nuclease digestion of nuclei reveals extended nucleosome ladders
914 having anomalous DNA lengths for chromatin assembled on non-replicating plasmids in
915 transfected cells. *Nucleic Acids Res* **22**: 370-375.
- 916 Jiang M, Anderson J, Gillespie J, Mayne M. 2008. uShuffle: a useful tool for shuffling biological sequences
917 while preserving the k-let counts. *BMC Bioinformatics* **9**:192.: 10.1186/1471-2105-1189-1192.
- 918 Kamiya H, Miyamoto S, Goto H, Kanda GN, Kobayashi M, Matsuoka I, Harashima H. 2013. Enhanced
919 transgene expression from chromatinized plasmid DNA in mouse liver. *Int J Pharm* **441**: 146-150.
920 doi: 110.1016/j.ijpharm.2012.1012.1004. Epub 2012 Dec 1012.
- 921 Kantor B, Ma H, Webster-Cyriaque J, Monahan PE, Kafri T. 2009. Epigenetic activation of unintegrated
922 HIV-1 genomes by gut-associated short chain fatty acids and its implications for HIV infection.
923 *Proc Natl Acad Sci U S A* **106**: 18786-18791. doi: 18710.11073/pnas.0905859106. Epub
924 0905852009 Oct 0905859120.
- 925 Kheradpour P, Ernst J, Melnikov A, Rogov P, Wang L, Zhang X, Alston J, Mikkelsen TS, Kellis M. 2013.
926 Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively
927 parallel reporter assay. *Genome Res* **23**: 800-811. doi: 810.1101/gr.144899.144112. Epub 142013
928 Mar 144819.
- 929 Kinney JB, Murugan A, Callan CG, Jr., Cox EC. 2010. Using deep sequencing to characterize the
930 biophysical mechanism of a transcriptional regulatory sequence. *Proc Natl Acad Sci U S A* **107**:
931 9158-9163. doi: 9110.1073/pnas.1004290107. Epub 1004292010 May 1004290103.
- 932 Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. 2014. A general framework for
933 estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46**: 310-315.
- 934 Kissler S, Stern P, Takahashi K, Hunter K, Peterson LB, Wicker LS. 2006. In vivo RNA interference
935 demonstrates a role for Nramp1 in modifying susceptibility to type 1 diabetes. *Nat Genet* **38**: 479-
936 483. Epub 2006 Mar 2019.
- 937 Klehr D, Maass K, Bode J. 1991. Scaffold-attached regions from the human interferon beta domain can be
938 used to enhance the stable expression of genes under the control of various promoters. *Biochemistry*
939 **30**: 1264-1270.
- 940 Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang
941 J, Ziller MJ et al. 2015. Integrative analysis of 111 reference human epigenomes. *Nature* **518**: 317-
942 330. doi: 310.1038/nature14248.
- 943 Kwaks TH, Barnett P, Hemrika W, Siersma T, Sewalt RG, Satijn DP, Brons JF, van Blokland R, Kwakman
944 P, Kruckeberg AL et al. 2003. Identification of anti-repressor elements that confer high and stable
945 protein production in mammalian cells. *Nat Biotechnol* **21**: 553-558. Epub 2003 Apr 2007.
- 946 Kwasniewski JC, Mogno I, Myers CA, Corbo JC, Cohen BA. 2012. Complex effects of nucleotide variants
947 in a mammalian cis-regulatory element. *Proc Natl Acad Sci U S A* **109**: 19498-19503. doi:
948 19410.11073/pnas.1210678109. Epub 1210672012 Nov 1210678105.
- 949 Leavitt AD, Robles G, Alesandro N, Varmus HE. 1996. Human immunodeficiency virus type 1 integrase
950 mutants retain in vitro integrase activity yet fail to integrate viral DNA efficiently during infection.
951 *J Virol* **70**: 721-728.

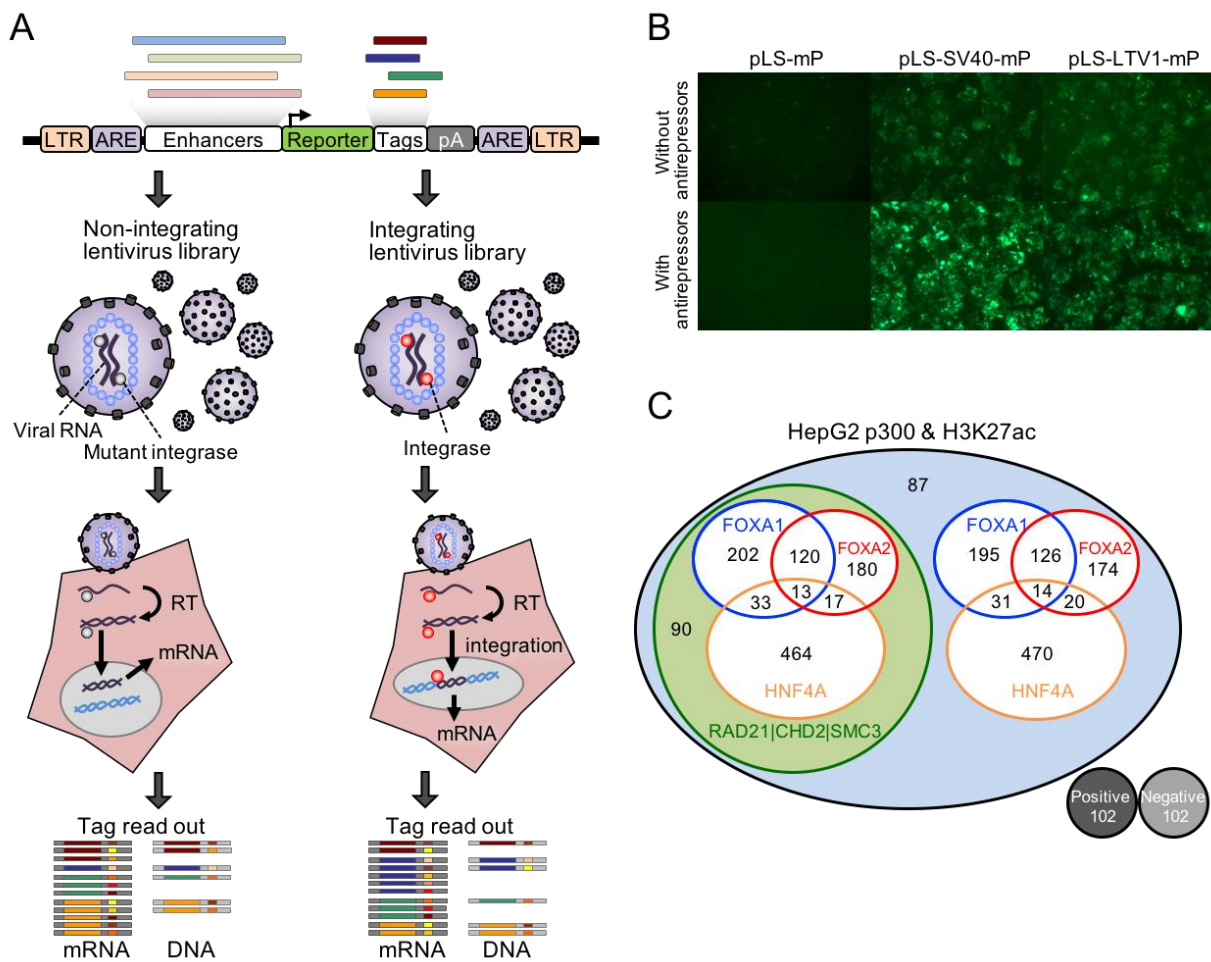
- 952 Lee D. 2016. LS-GKM: a new gkm-SVM for large-scale datasets. *Bioinformatics* **15**.
- 953 Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform.
- 954 *Bioinformatics* **25**: 1754-1760.
- 955 Lupien M, Eeckhoutte J, Meyer CA, Wang Q, Zhang Y, Li W, Carroll JS, Liu XS, Brown M. 2008. FoxA1
- 956 translates epigenetic signatures into enhancer-driven lineage-specific transcription. *Cell* **132**: 958-
- 957 970. doi: 910.1016/j.cell.2008.1001.1018.
- 958 Melnikov A, Murugan A, Zhang X, Tesileanu T, Wang L, Rogov P, Feizi S, Gnirke A, Callan CG, Jr.,
- 959 Kinney JB et al. 2012. Systematic dissection and optimization of inducible enhancers in human
- 960 cells using a massively parallel reporter assay. *Nat Biotechnol* **30**: 271-277. doi: 210.1038/nbt.2137.
- 961 Moreau P, Hen R, Wasylyk B, Everett R, Gaub MP, Chambon P. 1981. The SV40 72 base repair repeat has
- 962 a striking effect on gene expression both in SV40 and other chimeric recombinants. *Nucleic Acids*
- 963 *Res* **9**: 6047-6068.
- 964 Munir S, Thierry S, Subra F, Deprez E, Delelis O. 2013. Quantitative analysis of the time-course of viral
- 965 DNA forms during the HIV-1 life cycle. *Retrovirology* **10:87.**: 10.1186/1742-4690-1110-1187.
- 966 Murtha M, Tokcaer-Keskin Z, Tang Z, Strino F, Chen X, Wang Y, Xi X, Basilico C, Brown S, Bonneau R
- 967 et al. 2014. FIREWACH: high-throughput functional detection of transcriptional regulatory
- 968 modules in mammalian cells. *Nat Methods* **11**: 559-565. doi: 510.1038/nmeth.2885. Epub 2014
- 969 Mar 1023.
- 970 Nightingale SJ, Hollis RP, Pepper KA, Petersen D, Yu XJ, Yang C, Bahner I, Kohn DB. 2006. Transient
- 971 gene expression by nonintegrating lentiviral vectors. *Mol Ther* **13**: 1121-1132. Epub 2006 Mar
- 972 1123.
- 973 Patwardhan RP, Hiatt JB, Witten DM, Kim MJ, Smith RP, May D, Lee C, Andrie JM, Lee SI, Cooper GM
- 974 et al. 2012. Massively parallel functional dissection of mammalian enhancers in vivo. *Nat*
- 975 *Biotechnol* **30**: 265-270.
- 976 Patwardhan RP, Lee C, Litvin O, Young DL, Pe'er D, Shendure J. 2009. High-resolution analysis of DNA
- 977 regulatory elements by synthetic saturation mutagenesis. *Nat Biotechnol* **27**: 1173-1175.
- 978 Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features.
- 979 *Bioinformatics* **26**: 841-842.
- 980 Savic D, Roberts BS, Carleton JB, Partridge EC, White MA, Cohen BA, Cooper GM, Gertz J, Myers RM.
- 981 2015. Promoter-distal RNA polymerase II binding discriminates active from inactive CCAAT/
- 982 enhancer-binding protein beta binding sites. *Genome Res* **25**: 1791-1800. doi:
- 983 1710.1101/gr.191593.191115. Epub 192015 Oct 191520.
- 984 Schmidt D, Wilson MD, Ballester B, Schwalie PC, Brown GD, Marshall A, Kutter C, Watt S, Martinez-
- 985 Jimenez CP, Mackay S et al. 2010. Five-vertebrate ChIP-seq reveals the evolutionary dynamics of
- 986 transcription factor binding. *Science* **328**: 1036-1040. doi: 1010.1126/science.1186176. Epub
- 987 1182010 Apr 1186178.
- 988 Sharon E, van Dijk D, Kalma Y, Keren L, Manor O, Yakhini Z, Segal E. 2014. Probing the effect of
- 989 promoters on noise in gene expression using thousands of designed sequences. *Genome Res* **24**:
- 990 1698-1706. doi: 1610.1101/gr.168773.168113. Epub 162014 Jul 168716.
- 991 Shen SQ, Myers CA, Hughes AE, Byrne LC, Flannery JG, Corbo JC. 2016. Massively parallel cis-
- 992 regulatory analysis in the mammalian central nervous system. *Genome Res* **26**: 238-255. doi:
- 993 210.1101/gr.193789.193115. Epub 192015 Nov 193717.
- 994 Shlyueva D, Stampfel G, Stark A. 2014. Transcriptional enhancers: from properties to genome-wide
- 995 predictions. *Nat Rev Genet* **15**: 272-286. doi: 210.1038/nrg3682. Epub 2014 Mar 1011.
- 996 Sloan CA, Chan ET, Davidson JM, Malladi VS, Strattan JS, Hitz BC, Gabdank I, Narayanan AK, Ho M,
- 997 Lee BT et al. 2016. ENCODE data at the ENCODE portal. *Nucleic Acids Res* **44**: D726-732. doi:
- 998 710.1093/nar/gkv1160. Epub 2015 Nov 1092.
- 999 Smith CL, Hager GL. 1997. Transcriptional regulation of mammalian genes in vivo. A tale of two
- 1000 templates. *J Biol Chem* **272**: 27493-27496.

- 1001 Smith RP, Taher L, Patwardhan RP, Kim MJ, Inoue F, Shendure J, Ovcharenko I, Ahituv N. 2013.
1002 Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational
1003 model. *Nat Genet* **45**: 1021-1028.
- 1004 Thierry S, Thierry E, Subra F, Deprez E, Leh H, Bury-Mone S, Delelis O. 2016. Opposite transcriptional
1005 regulation of integrated vs unintegrated HIV genomes by the NF-kappaB pathway. *Sci Rep*
1006 **6:25678.**: 10.1038/srep25678.
- 1007 Tsukada J, Yoshida Y, Kominato Y, Auron PE. 2011. The CCAAT/enhancer (C/EBP) family of basic-
1008 leucine zipper (bZIP) transcription factors is a multifaceted highly-regulated system for gene
1009 regulation. *Cytokine* **54**: 6-19. doi: 10.1016/j.cyto.2010.1012.1019. Epub 2011 Jan 1022.
- 1010 Visel A, Blow MJ, Li Z, Zhang T, Akiyama JA, Holt A, Plajzer-Frick I, Shoukry M, Wright C, Chen F et
1011 al. 2009. ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* **457**: 854-858.
- 1012 Wang X, McManus M. 2009. Lentivirus production. *J Vis Exp* (**32**). 1499. doi: 1410.3791/1499.
- 1013 Watt AJ, Garrison WD, Duncan SA. 2003. HNF4: a central regulator of hepatocyte differentiation and
1014 function. *Hepatology* **37**: 1249-1253.
- 1015 White MA. 2015. Understanding how cis-regulatory function is encoded in DNA sequence using massively
1016 parallel reporter assays and designed sequences. *Genomics* **106**: 165-170. doi:
1017 110.1016/j.ygeno.2015.1006.1003. Epub 2015 Jun 1010.
- 1018 Zerbino DR, Wilder SP, Johnson N, Juettemann T, Flicek PR. 2015. The ensembl regulatory build. *Genome*
1019 *Biol* **16:56.**: 10.1186/s13059-13015-10621-13055.
- 1020
- 1021

Figures

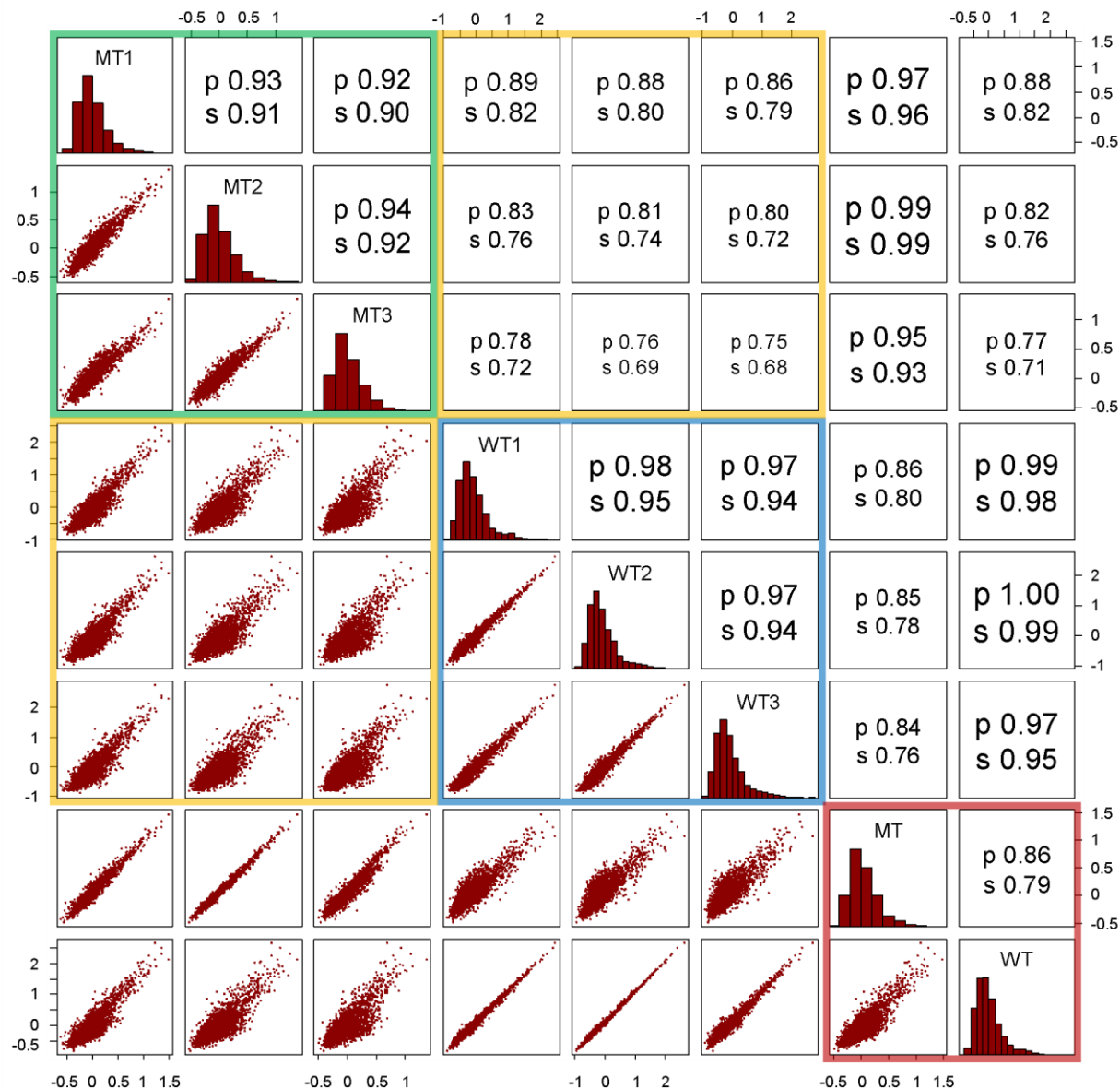
1022
1023
1024
1025
1026
1027
1028
1029
1030
1031
1032
1033
1034
1035
1036
1037

Figure 1. Study design for LentiMPRA. (A) Schematic diagram of lentiMPRA. Candidate enhancers and barcode tags were synthesized in tandem as a microarray-derived oligonucleotide library, and cloned into the pLS-mP vector, followed by cloning of a minimal promoter (mP) and reporter (EGFP) between them. The resulting lentiMPRA library was packaged with either wildtype or mutant integrase, and infected into HepG2 cells. Both DNA and mRNA were extracted, and barcode tags were sequenced to test their enhancer activities in an episomal vs. genome integrating manner. (B) HepG2 cells infected with lentiviral reporter construct bearing no enhancer (pLS-mP), an SV40 enhancer (pLS-SV40-mP), and LTV1 (pLS-LTV1-mP), a known liver enhancer (Patwardhan et al. 2012), with or without antirepressors. The inclusion of antirepressors results in stronger and more consistent expression, but is still dependent on the presence of an enhancer. (C) Venn diagram showing the composition of the lentiMPRA library. 2,236 enhancer candidate sequences were chosen on the basis of having ENCODE HepG2 ChIP-seq peaks for EP300 and H3K27ac marks. The candidates overlapped with or without ChIP-seq peaks for FOXA1, FOXA2, or HNF4A. Half of the candidates overlapped with ChIP-seq peaks for RAD21, SMC3, and CHD2. The library included 102 positive and 102 negative controls.



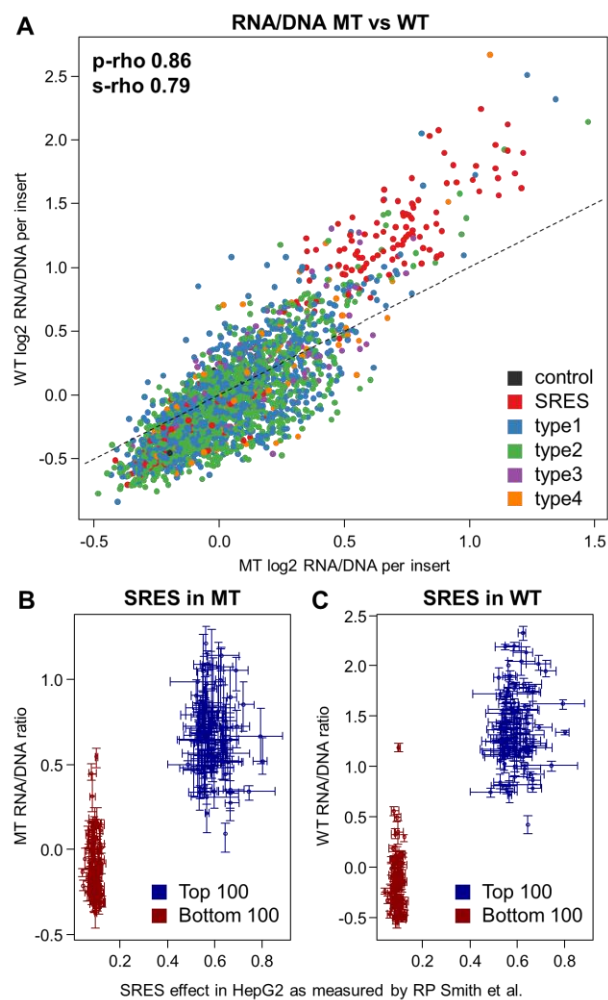
1038

1039 **Figure 2. Pairwise correlation of per-insert RNA/DNA ratios between replicates, within and between**
 1040 **MT vs. WT experiments.** The lower left triangle shows pair-wise scatter plots. The diagonal provides
 1041 replicate names and the respective histogram of the RNA/DNA ratios for that replicate. The upper triangle
 1042 provides Pearson (p) and Spearman (s) correlation coefficients. MT vs. MT (green box) or WT vs. WT
 1043 (blue box) comparisons are substantially more correlated than MT vs. WT (yellow boxes) comparisons,
 1044 consistent with systematic differences between the episomal vs. integrated contexts for reporter assays that
 1045 exceed technical noise. The two right-most columns and two bottom-most rows correspond to MT and WT
 1046 after combining across the three replicates, with the combined MT vs. the combined WT comparison in the
 1047 red box.



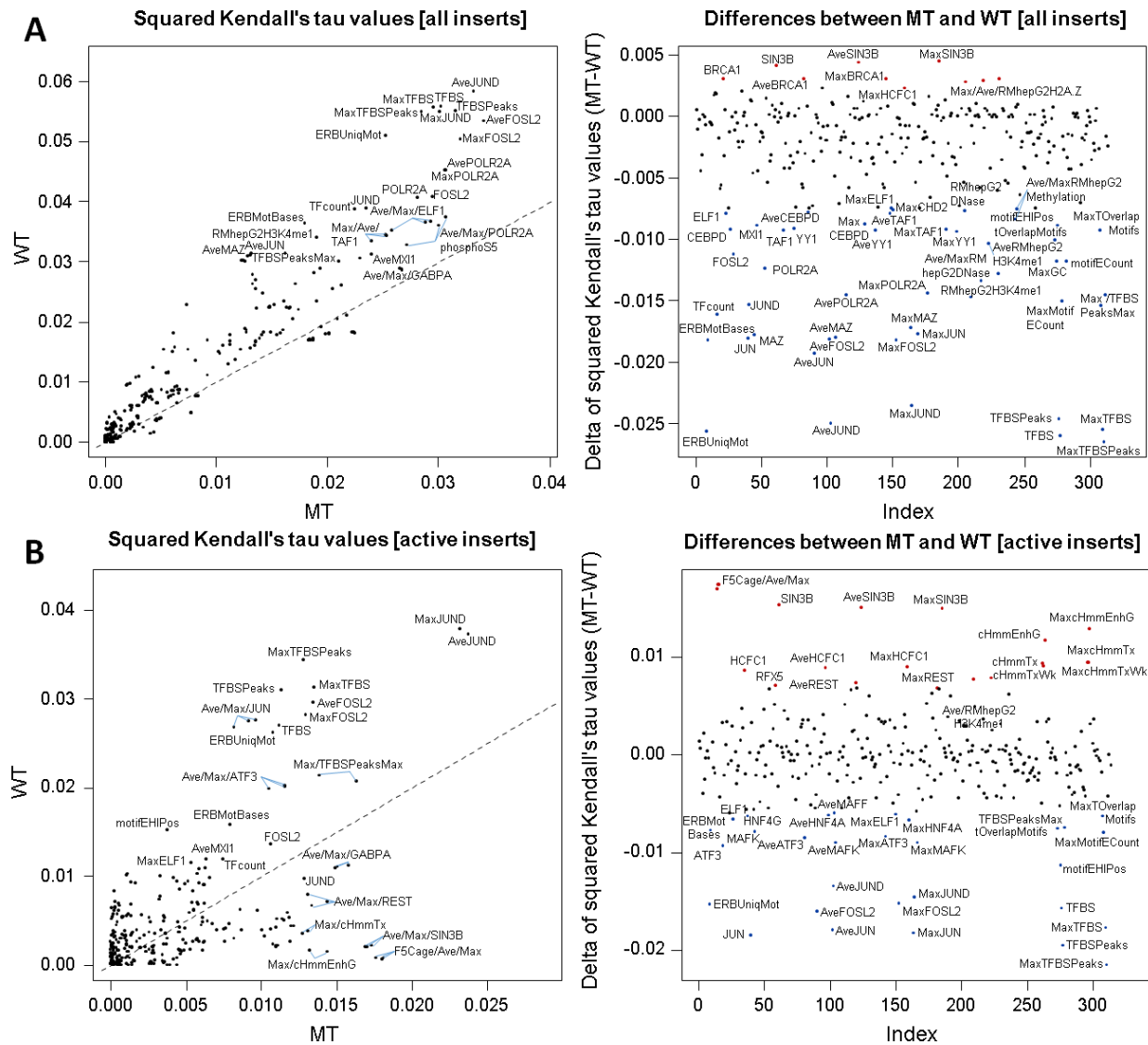
1048
 1049

1050 **Figure 3. Comparisons between the non-integrating (MT) and integrating (WT) libraries** (A) Scatter
1051 plot of combined MT vs. WT RNA/DNA ratios. MT ratios show a smaller dynamic range and thus seem
1052 compressed compared to WT results. Data points are colored by the type of insert sequence, including two
1053 types of controls: a total of four positive and negative controls (black) as well as the highest 100 and lowest
1054 100 synthetic regulatory element sequences (SRES, red) identified by Smith *et al.* (Smith et al. 2013). The
1055 four classes of putative enhancer elements are: Regions of FOXA1, FOXA2 or HNF4A binding that overlap
1056 H3K27ac and EP300 calls as well as at least one of three chromatin remodeling factors RAD21, CHD2 or
1057 SMC3 (type 1); Regions like in 1 but with no remodeling factor overlapping (type 2); EP300 peak regions
1058 overlapping H3K27ac as well as at least one of chromatin remodeling factor, but without peaks in FOXA1,
1059 FOXA2 or HNF4A (type 3); Regions like in 3 but with no remodeling factor overlapping (type 4). As
1060 shown here and in Fig. S11, we do not observe major differences between the four design types, either with
1061 respect to activity or MT vs. WT. (B+C) Enhancer activity of 200 synthetic regulatory element sequences
1062 (SRES) in the MT (B) and WT experiments (C). Scatter plot of RNA/DNA ratios for the top 100 positive
1063 and top 100 negative synthetic regulatory element (SRE) sequences in HepG2 experiments by Smith *et al.*
1064 (Smith et al. 2013). Plots show the combined RNA/DNA ratios on the y-axis and measurements by Smith
1065 *et al.* on the x-axis. Intervals indicate the mean, minimum and maximum values observed for three replicates
1066 performed with each experiment.



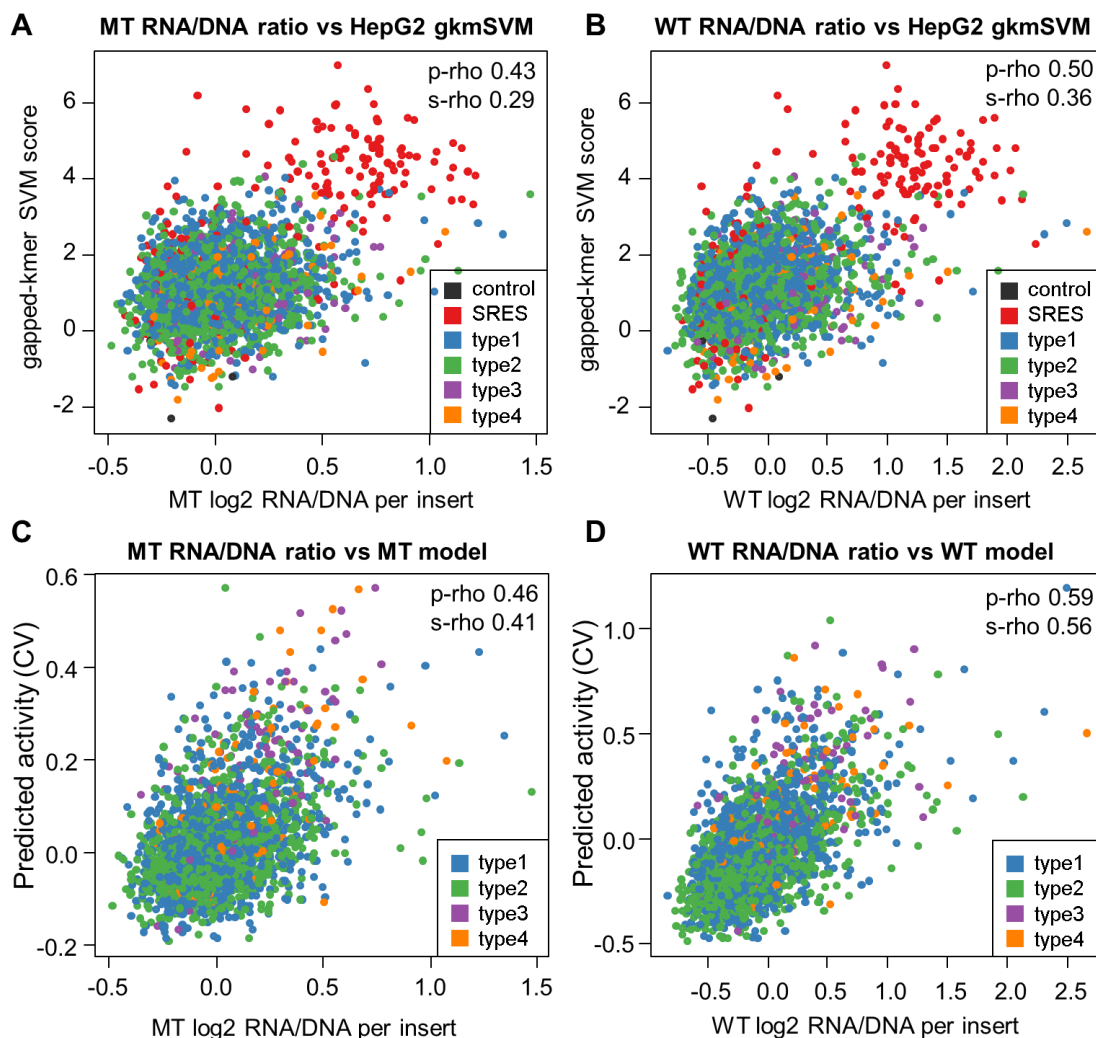
1067

1068 **Figure 4. Squared Kendall's tau (T^2) values for available genome annotations for predicting the**
 1069 **activity of candidate enhancer sequences in the non-integrating (MT) and integrating (WT)**
 1070 **experiments. (A) WT RNA/DNA ratios correlate better with annotations than the respective MT values.**
 1071 **The left panel highlights the top correlated annotations for WT and MT ratios. The right panel highlights**
 1072 **annotations with the largest difference in T^2 values between the MT and WT experiments. (B) Same analysis**
 1073 **for the subset of active elements (Suppl. Table 1).**



1074
1075

1076 **Figure 5. Prediction models.** (A+B) Correlation of gkm-SVM scores obtained for a combined HepG2
1077 model with RNA/DNA ratios obtained from the mutant (MT) and wild-type integrase (WT) experiments.
1078 Data points are colored by the type of insert sequence, including two types of controls, 200 synthetic
1079 regulatory element sequences (SRES, red) identified by Smith *et al.* (Smith et al. 2013) and four other
1080 control sequences (dark gray). The four classes of putative enhancer elements are: (type 1) regions of
1081 FOXA1, FOXA2 or HNF4A binding that overlap H3K27ac and EP300 calls as well as at least one of three
1082 chromatin remodeling factors RAD21, CHD2 or SMC3; (type 2) regions like in 1 but with no remodeling
1083 factor overlapping; (type 3) EP300 peak regions overlapping H3K27ac as well as at least one of chromatin
1084 remodeling factor, but without peaks in FOXA1, FOXA2 or HNF4A; (type 4) regions like in 3 but with no
1085 remodeling factor overlapping. Correlations are partially driven by the SRES; when excluding all controls,
1086 Spearman's R^2 values drop from 0.0817 to 0.0409 and from 0.1282 to 0.0756 for MT and WT, respectively.
1087 (C+D) Scatter plots of measured RNA/DNA ratios with predicted activity from linear Lasso models using
1088 annotations (numerical and categorical) as well as sequence-based (individual LS-GKM scores)
1089 information. Correlation coefficients are 0.46 Pearson / 0.41 Spearman for the non-integrated experiment
1090 (MT) and 0.59 Pearson / 0.56 Spearman for the integrated constructs (WT). The models selected 111 (MT)
1091 and 135 (WT) out of a total of 378 annotation features. Based on Pearson R^2 values, these combined models
1092 explain 21.1% (MT) and 34.7% (WT) of the variance observed in these experiments.

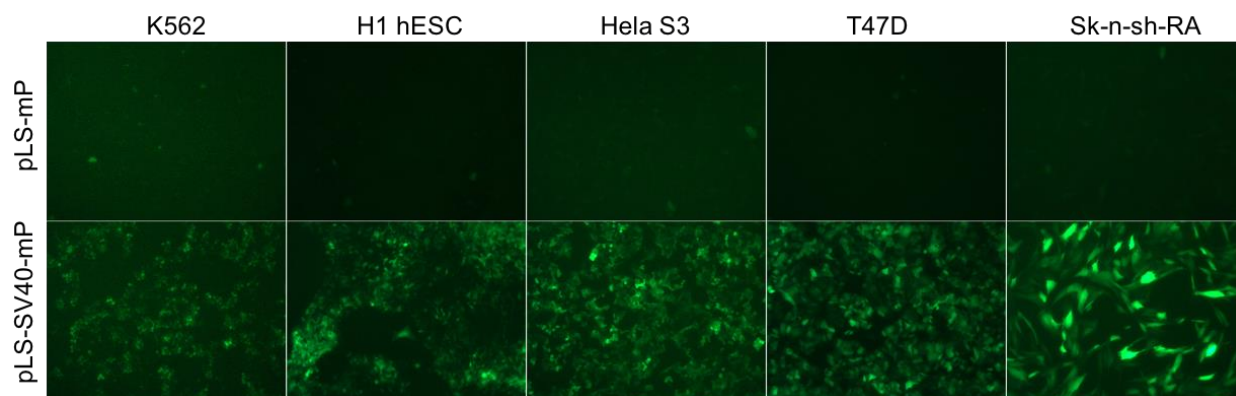


1093

1094
1095
1096
1097
1098

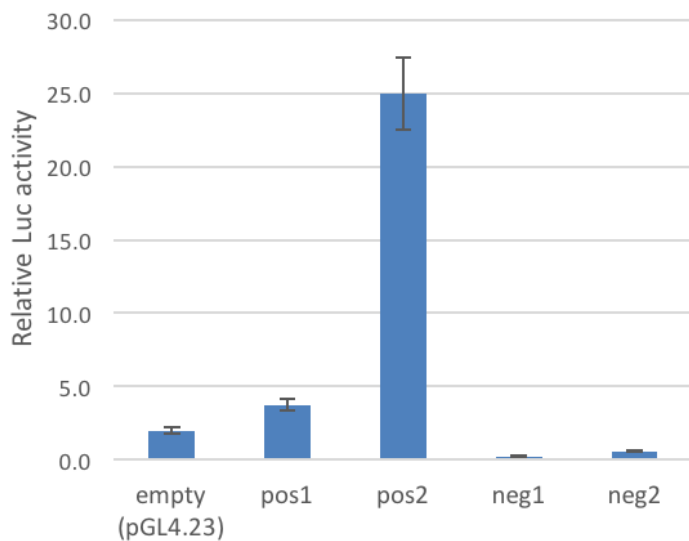
Supplementary figures

Supplemental Figure S1. Validating lentiviral infection in several different cell lines. pLS-mP or pLS-SV40-mP infection results following 48 hours for K562, H1 human embryonic stem cells (H1 hESC), HeLa S3, T47D, and Sk-n-sh cells treated with retinoic acid (Sk-n-sh-RA).



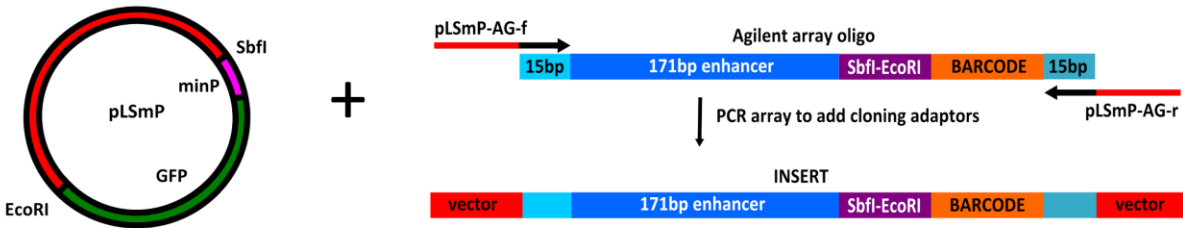
1099
1100

1101 **Supplemental Figure S2.** Luciferase assay for two positive (pos1 and pos2) and two negative (neg1 and
1102 neg2) control sequences. The reporter vectors and the empty vector (pGL4.23) were transfected into HepG2
1103 cells, and the enhancer activity was measured as relative luciferase activity compared to Renilla luciferase
1104 expression following 24 hours.

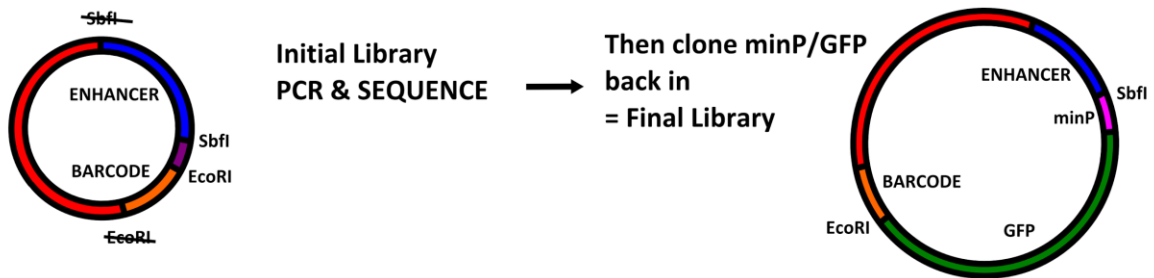


1105
1106

1107 **Supplemental Figure S3.** MPRA library cloning strategy. GFP, green fluorescent protein, minP, minimal
1108 promoter; *SbfI*, *EcoRI* restriction sites. The lenti vector pLS-mP is cut with *SbfI* and *EcoRI* to remove the
1109 minimal promoter and GFP. The enhancer/barcode agilent array is amplified with adaptor primers, and
1110 these PCR products are cloned into the pLS-mP backbone using NEBuilder HiFi Assembly mix. Cloning
1111 disrupts the original *SbfI* and *EcoRI* sites. This initial library can be sequenced to validate its accuracy and
1112 complexity. The library is then digested with *SbfI* and *EcoRI* to re-insert the minimal promoter and GFP
1113 between the enhancer and the barcode with a sticky-end ligation.

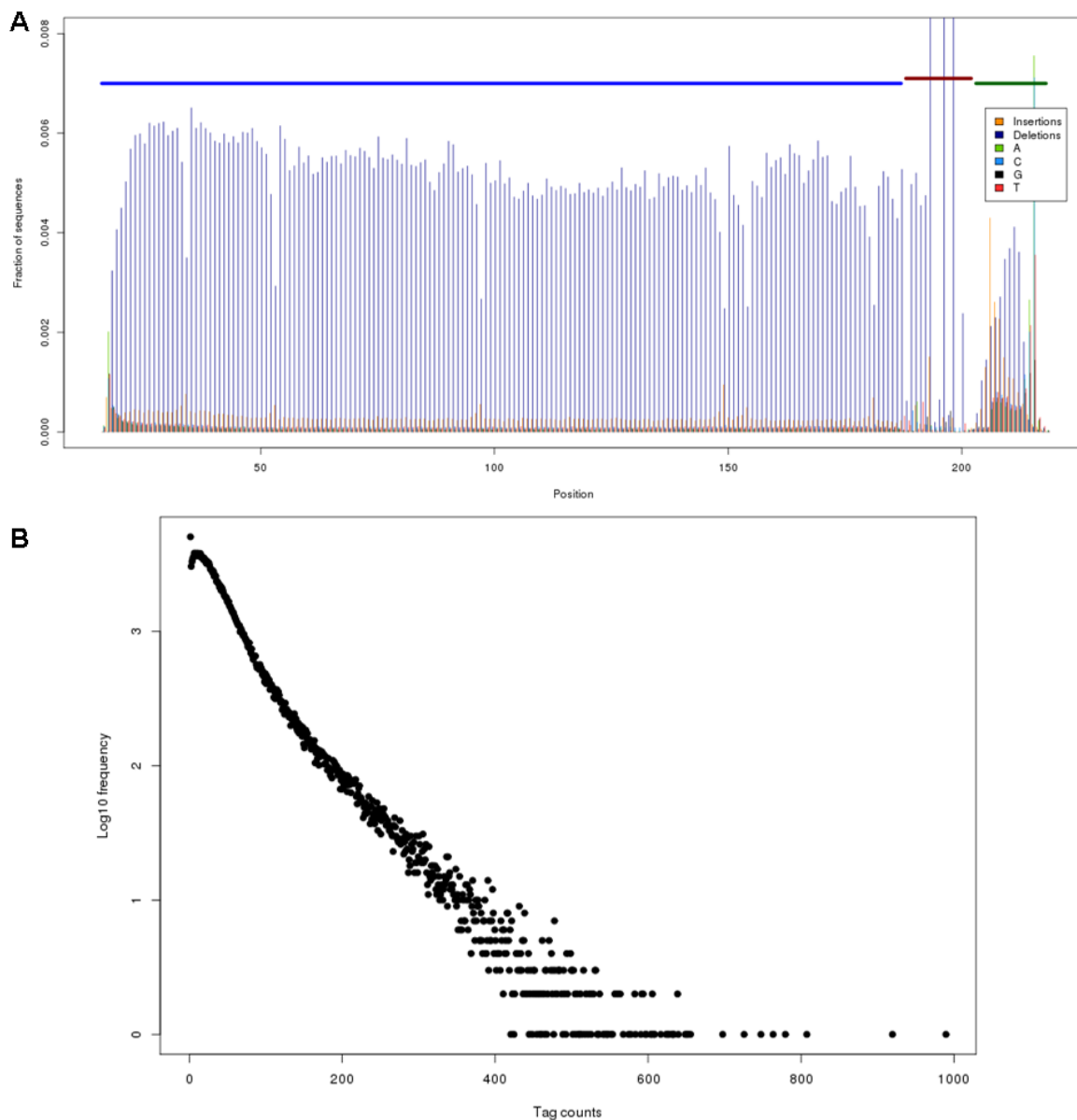


Cut pLSmP with *SbfI* and *EcoRI*, remove minP/GFP and replace with enhancer/barcode inserts



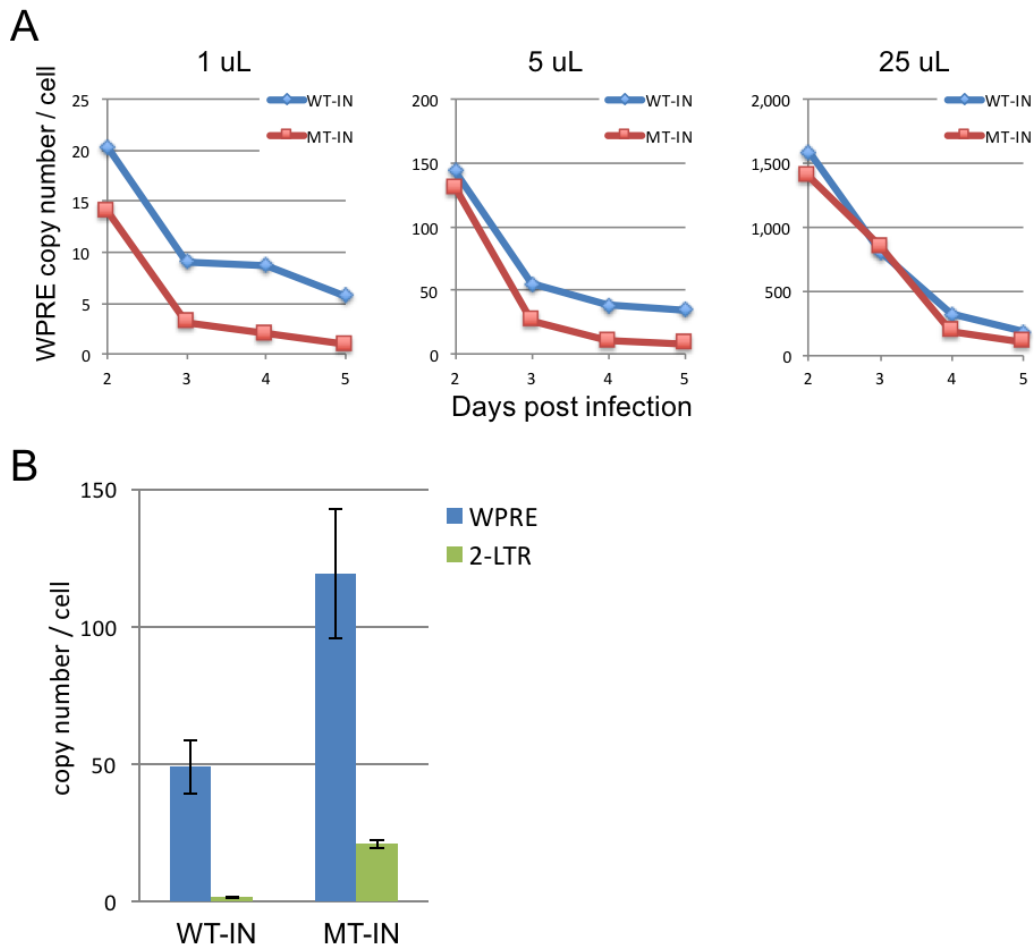
1114
1115
1116
1117

1118 **Supplemental Figure S4.** Sequence quality and barcode representation in the array-synthesized
1119 lentiMPRA library. (A) Positioning of errors/differences to the designed oligo sequences as observed from
1120 consensus-called BWA alignments. Thick horizontal lines near the top indicate the different oligo parts
1121 (insert - blue, spacer - red, barcode - green). Deletions towards read ends are misrepresented in this plot as
1122 those that result in altered outer alignment coordinates rather than the identification of a deletion. Filtering
1123 for sequence mapability results in artifacts around the barcode sequence which is required to disambiguate
1124 oligos. (B) Number of times designed reporter barcodes are observed in the oligo validation sequencing
1125 experiment. The frequency axis (y) has been log-transformed to show an over dispersion effect in the
1126 library, where a minority of barcodes contribute many of the observations. This effect can be observed from
1127 a change in gradient between barcodes observed below 150 times and barcodes observed more than 150
1128 times. The 11,889 barcodes (5.5%) observed with more than 150 observations account for 24.5% of all
1129 observations.
1130



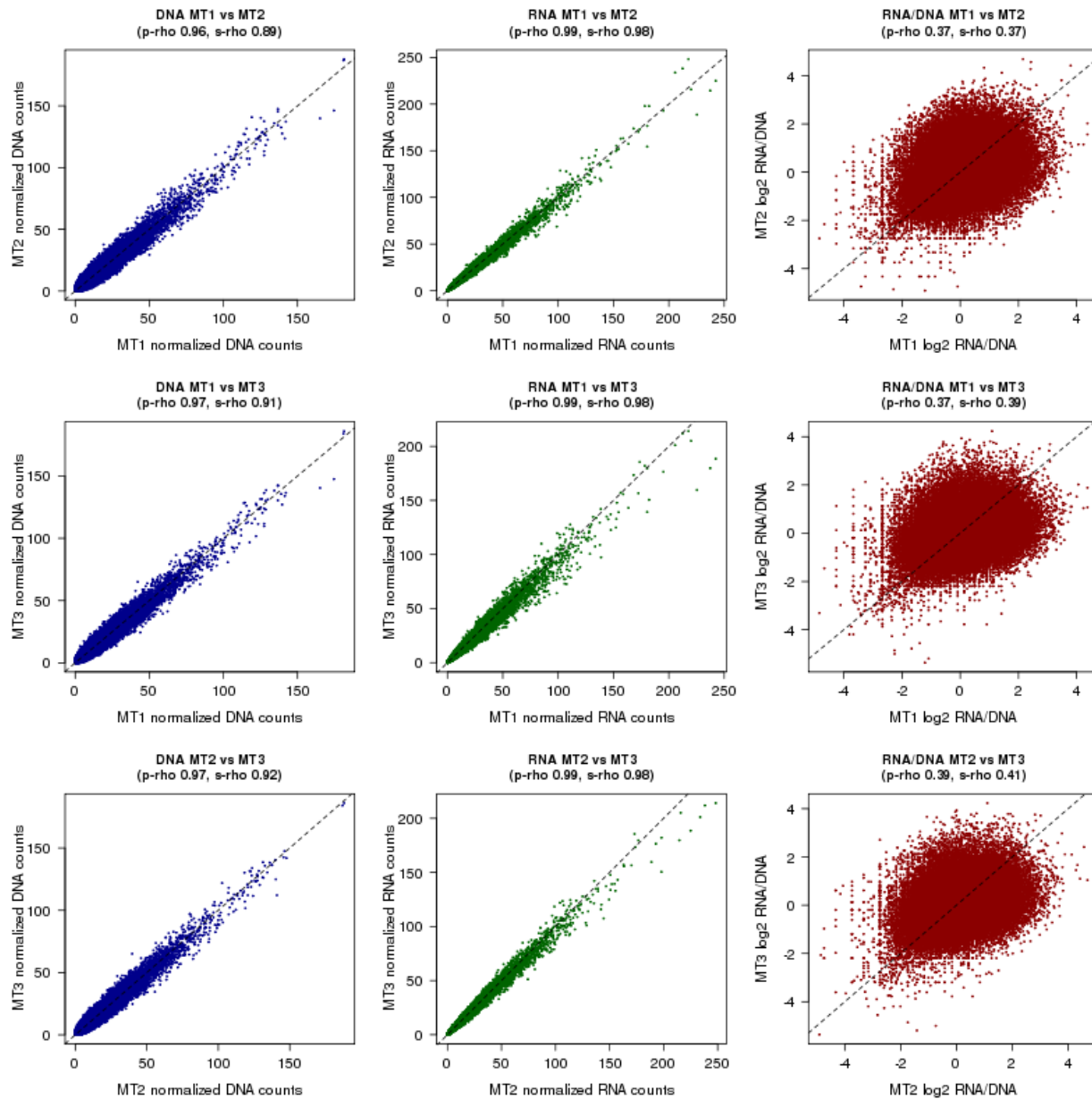
1131

1132 **Supplemental Figure S5.** Analysis of wild-type integrase (WT-IN) and mutant integrase (MT-IN) virus
1133 infection conditions. (A) DNA titer for pLS-SV40-mP using either a WT-IN and MT-IN virus as determined
1134 by qPCR with primers against WPRE compared to genomic primers for the intronic region of the *LIPC*
1135 gene. Three different volumes (1, 5, and 25 μ L per well of a 24-well plate) of lentivirus were analyzed at
1136 days 2-5 days. (B) Analysis of total viral DNA (WPRE) and unintegrated viral DNA (2-LTR circular DNA)
1137 for WT-IN library 4 days after infection and MT-IN library 3 days after infection using qPCR. The relative
1138 amount of viral DNA was measured using primers that target an intronic region of the *LIPC* gene.
1139



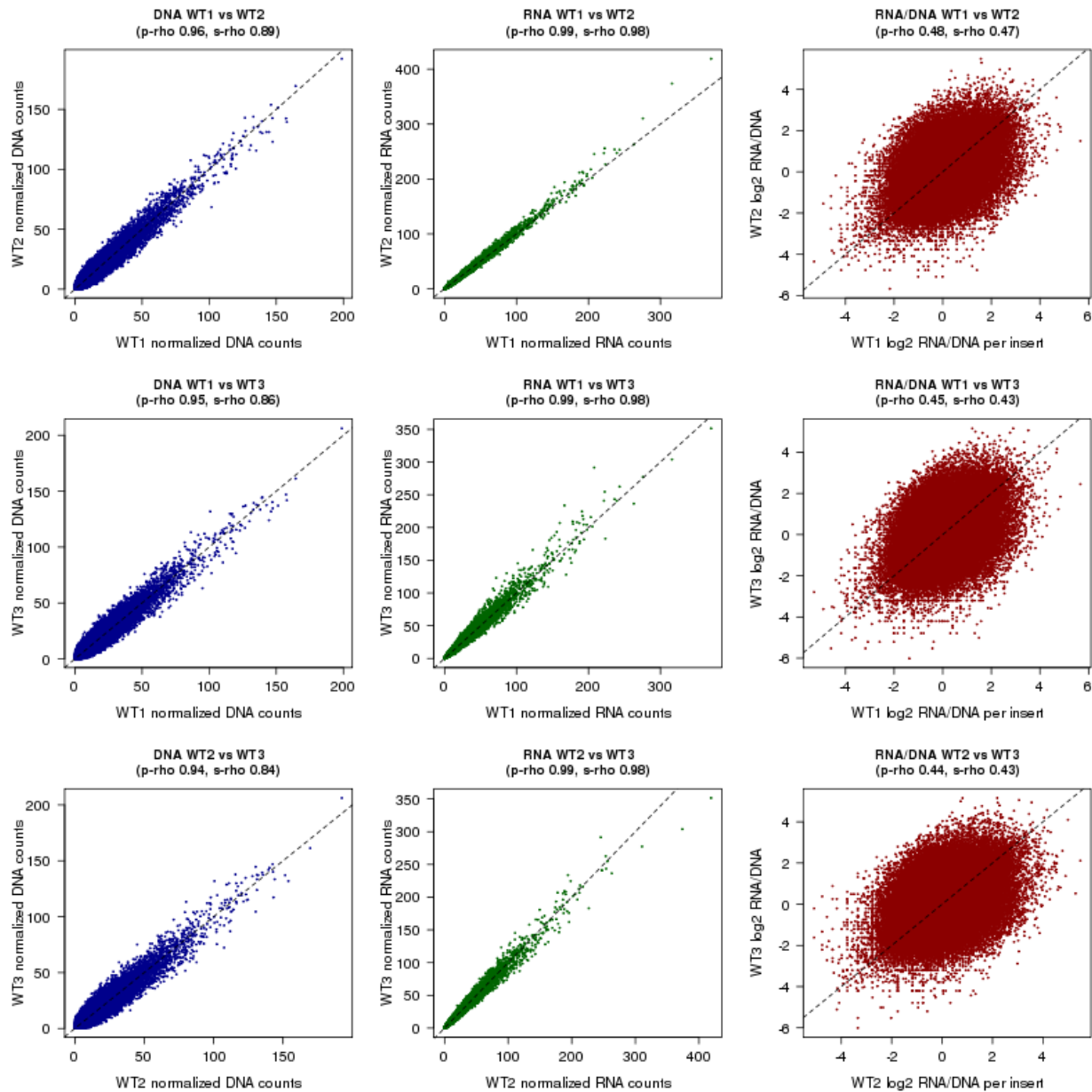
1140

1141 **Supplemental Figure S6.** Correlation (Pearson/p-rho and Spearman/s-rho) of DNA tag counts (left
1142 column), RNA tag counts (middle column) as well as RNA/DNA ratios (right column) for pairwise
1143 comparisons of three mutant-integrase (MT) library replicates (rows). These are for individual barcodes,
1144 *i.e.* before summing across barcodes representing the same candidate enhancer sequence.
1145



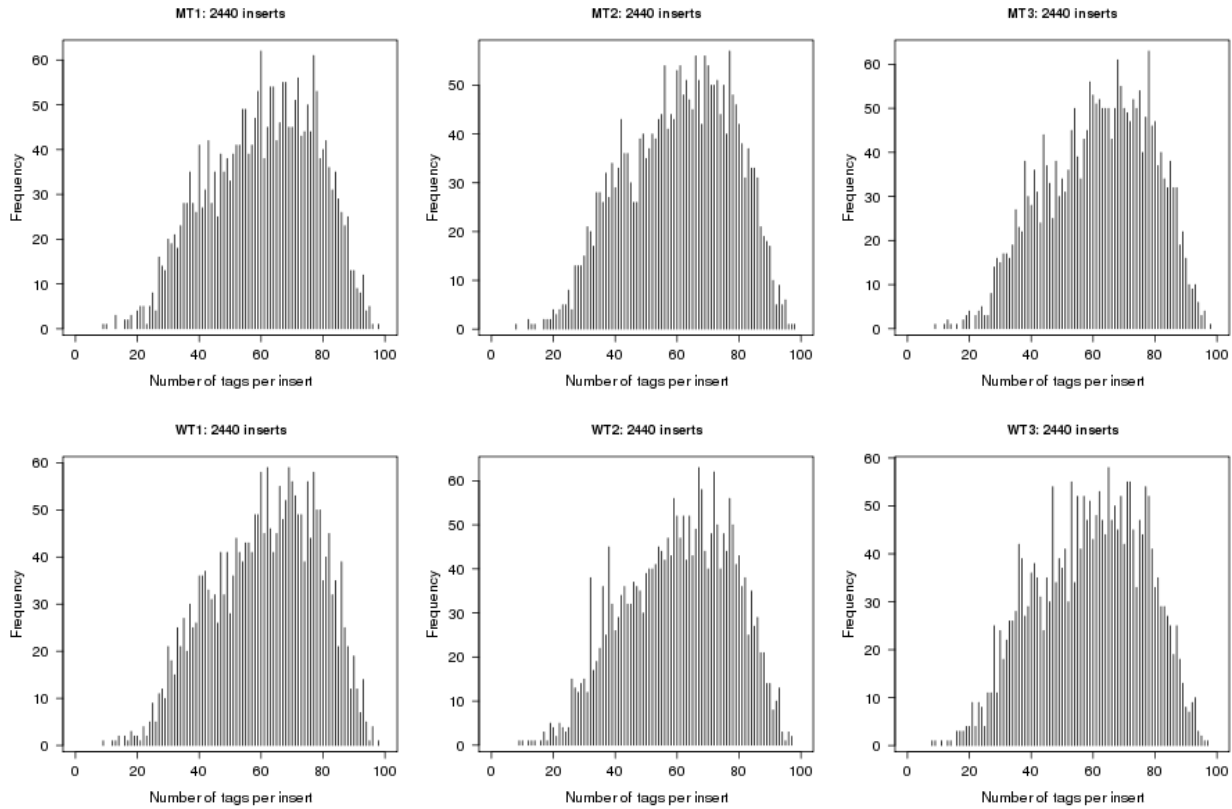
1146

1147 **Supplemental Figure S7.** Correlation (Pearson/p-rho and Spearman/s-rho) of DNA tag counts (left
1148 column), RNA tag counts (middle column) as well as RNA/DNA ratios (right column) for pairwise
1149 comparisons of three wild-type-integrase (WT) library replicates (rows). These are for individual barcodes,
1150 *i.e.* before summing across barcodes representing the same candidate enhancer sequence.
1151



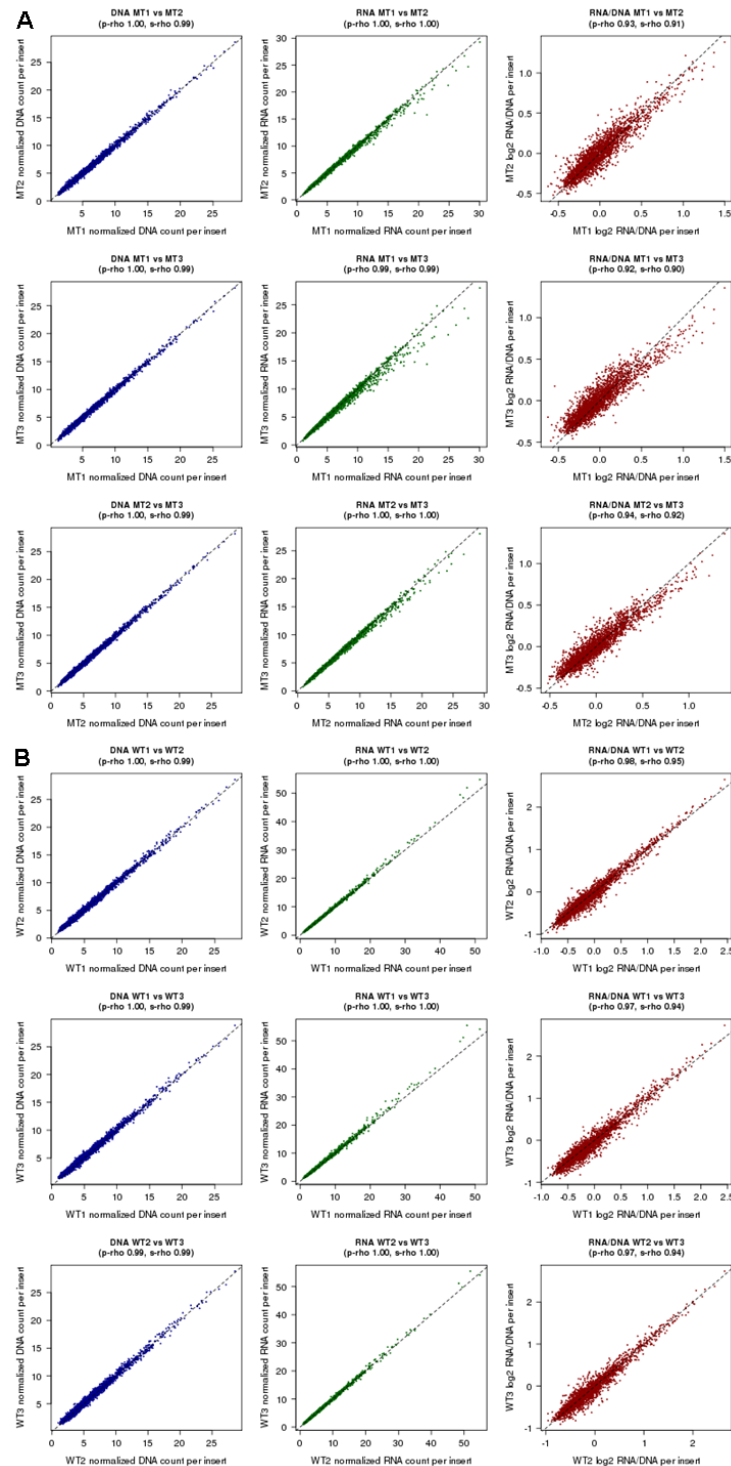
1152

1153 **Supplemental Figure S8.** Histogram of the number of barcodes per insert across all 2,440 designed
1154 candidate enhancer sequences. We observe barcodes for all 2,440 inserts in each of the six experiments. On
1155 average, we observe 59-62 barcodes per insert (minimum 8-9, maximum 97-98).
1156



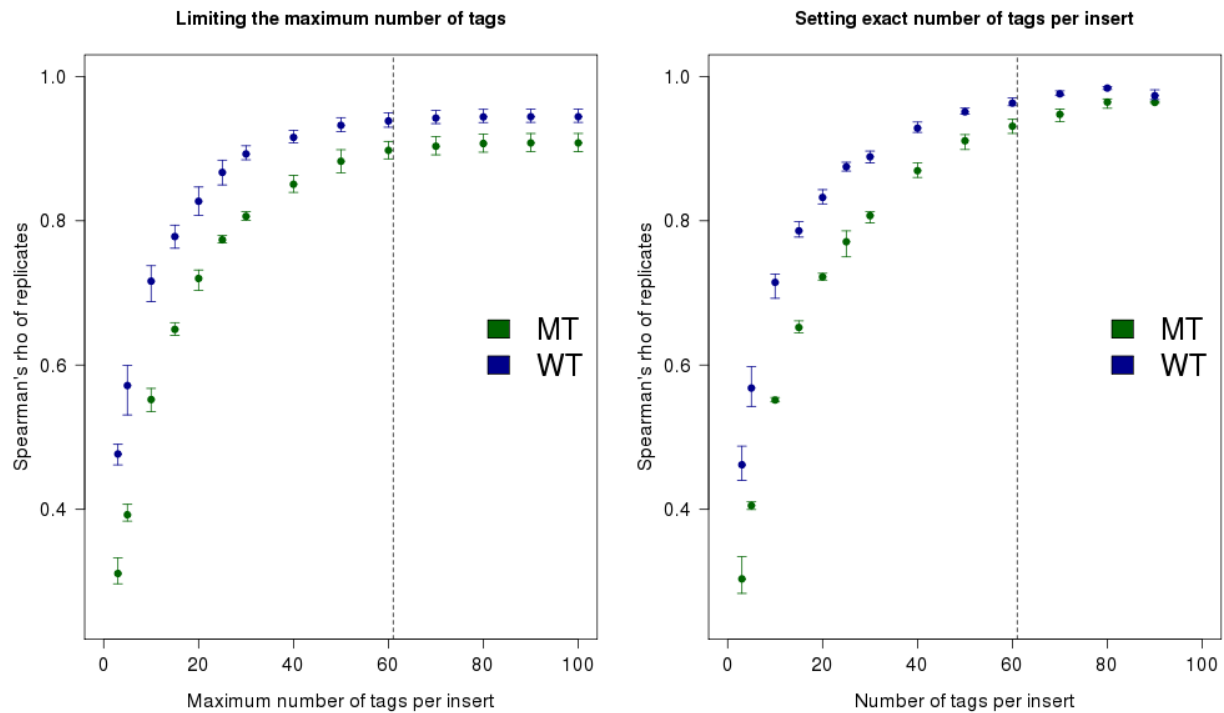
1157

1158 **Supplemental Figure 9. Reproducibility of counts and ratios across technical replicates.** Correlation
1159 of DNA tag counts (left/blue), RNA tag counts (middle/green) as well as RNA/DNA ratios (right/red) for
1160 pairwise comparisons of the three mutant-integrase (MT) library replicates (A) and the three wild-type-
1161 integrase (WT) library replicates (B) along the rows. Both the counts and the ratios are calculated from
1162 RNA and DNA counts that are summed across ~60 barcodes representing a given candidate enhancer
1163 sequence and represented in a given experiment.
1164



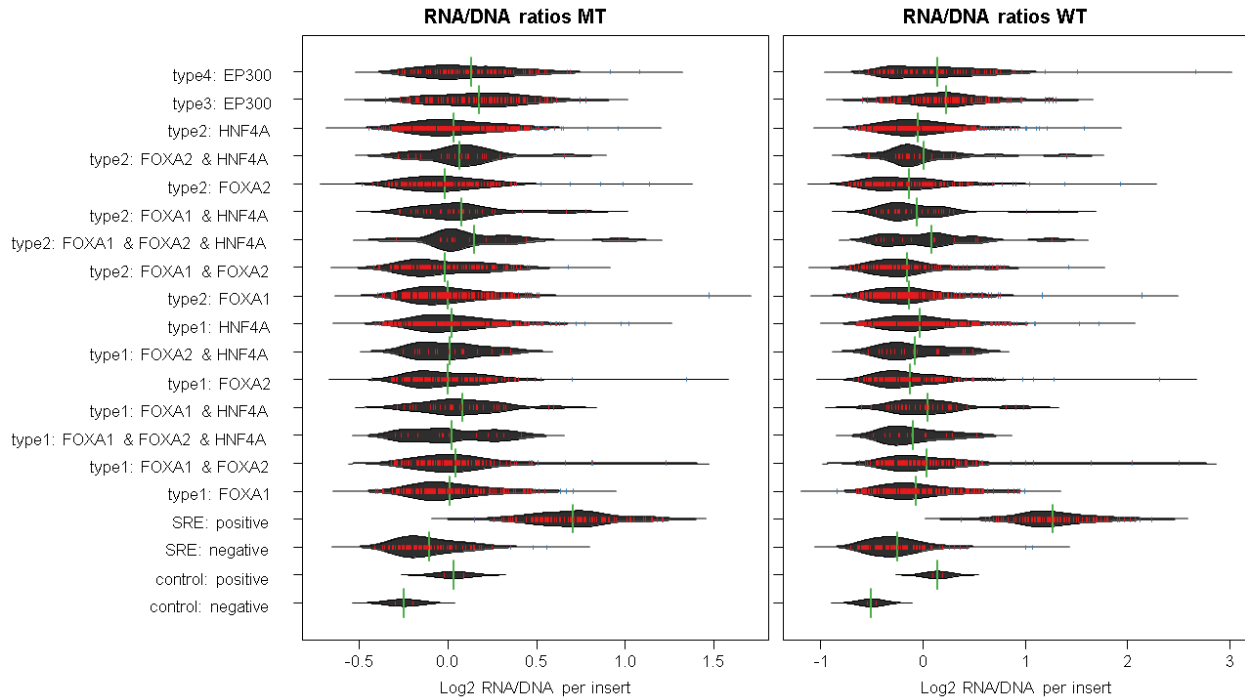
1165

1166 **Supplemental Figure S10.** Effect of the number of barcodes averaged per insert on the correlation of
1167 experimental replicates for mutant (MT) and wild-type integrase (WT) experiments. Intervals indicate the
1168 mean, minimum and maximum values observed for pairwise Spearman correlations of the three replicates.
1169 Dashed lines indicate the average number of barcodes per insert as observed across all six experiments. The
1170 left plot is with setting a maximum number of tags per insert. The right plot is with fixing an exact number
1171 of tags per insert.
1172



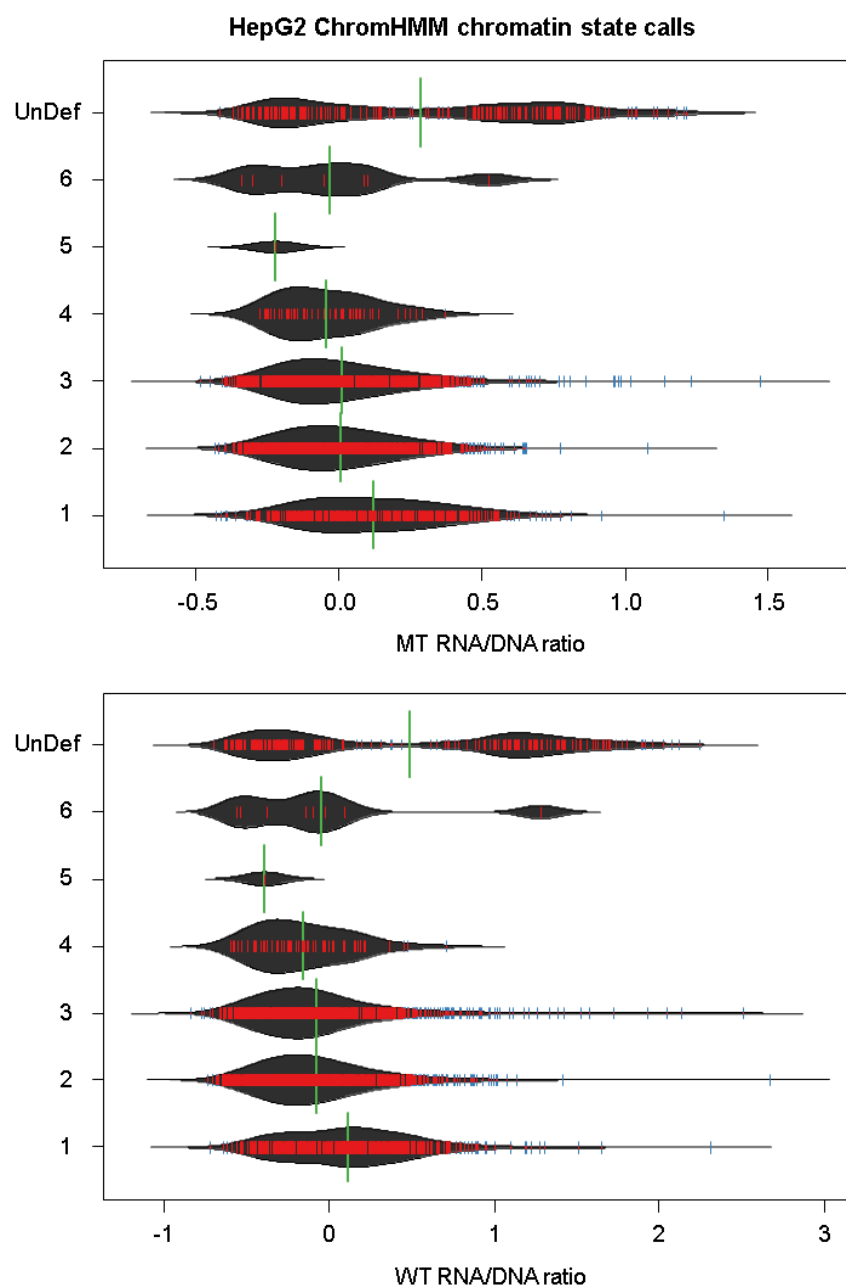
1173

1174 **Supplemental Figure S11.** Bean plot of MT (left) and WT (right) RNA/DNA ratios split by design
1175 category. The four classes of putative enhancer elements are: Regions of FOXA1, FOXA2 or HNF4A
1176 binding that overlap H3K27ac and EP300 calls as well as at least one of three chromatin remodeling factors
1177 RAD21, CHD2 or SMC3 (type 1); Regions like in 1 but with no remodeling factor overlapping (type 2);
1178 EP300 peak regions overlapping H3K27ac as well as at least one of chromatin remodeling factor, but
1179 without peaks in FOXA1, FOXA2 or HNF4A (type 3); Regions like in 3 but with no remodeling factor
1180 overlapping (type 4). As shown here and in Fig. 3A, we do not observe major differences between design
1181 types, either with respect to activity or MT vs. WT.
1182



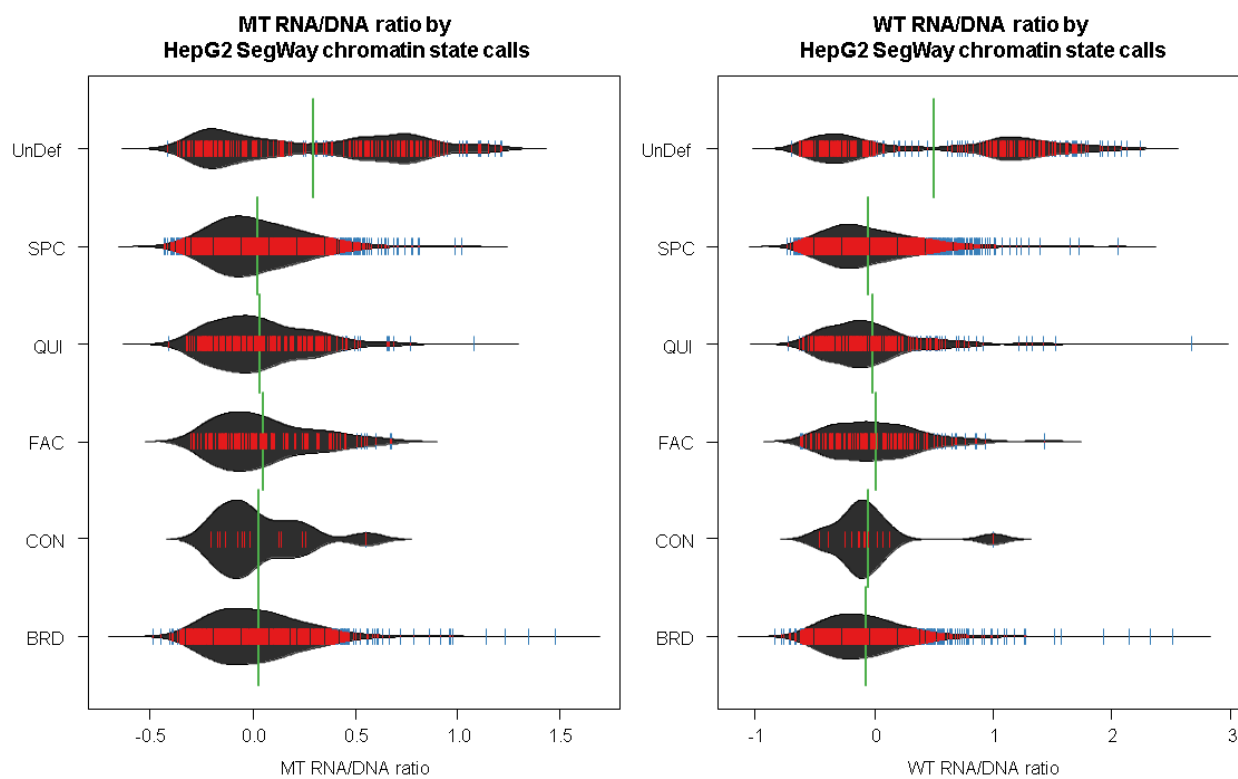
1183

1184 **Supplemental Figure S12.** Distribution of RNA/DNA ratios for MT (top) and WT (bottom) split by HepG2
1185 ChromHMM states. ChromHMM states were downloaded from the ENCODE project and had not been
1186 annotated with further labels. Inserts not represented in the available annotations were assigned to UnDef.
1187



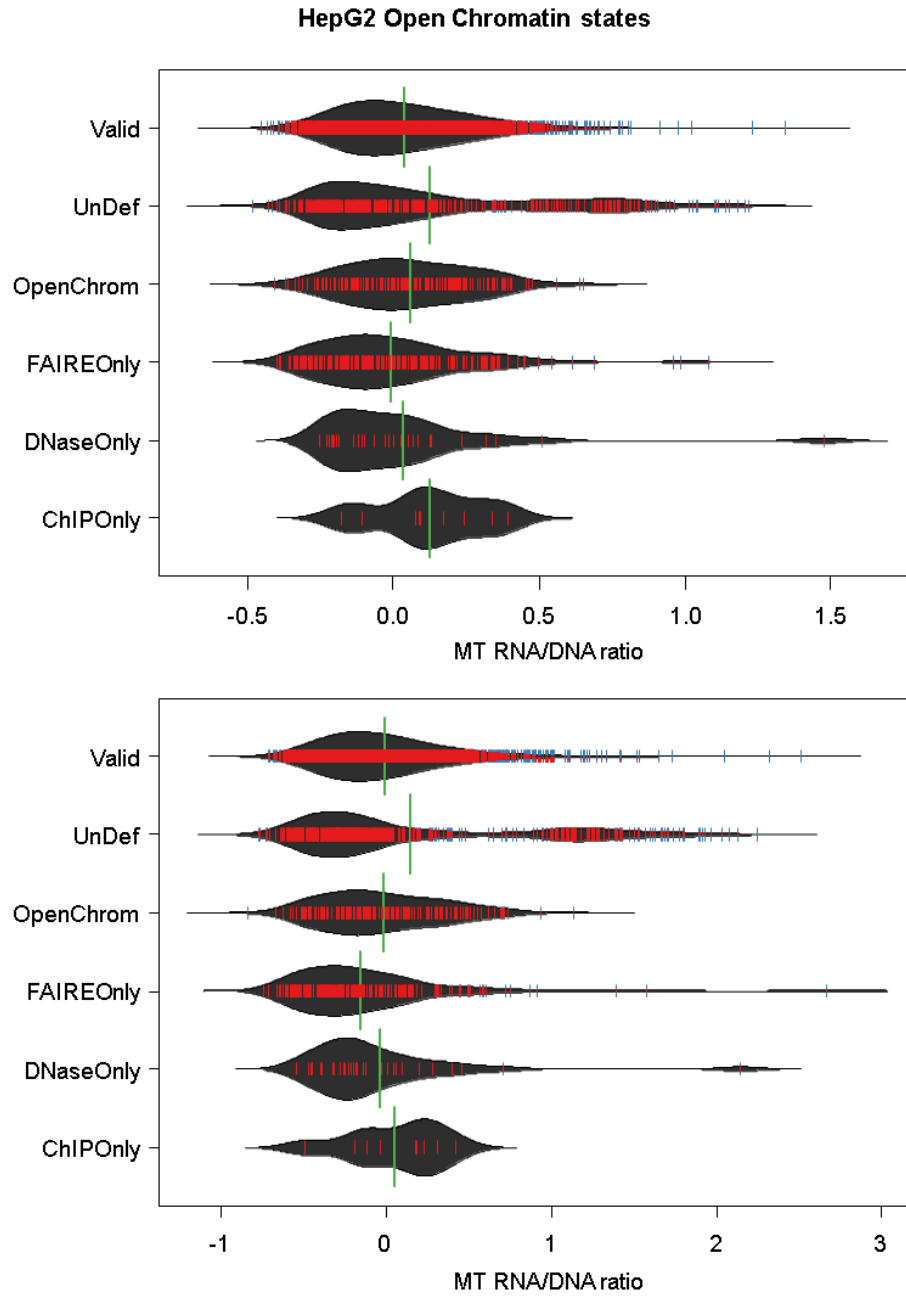
1188

1189 **Supplemental Figure S13.** Distribution of RNA/DNA ratios for MT (left) and WT (right) split by HepG2
1190 SegWay chromatin states. Chromatin labels were downloaded from the ENCODE project. Inserts not
1191 represented in the available annotations were assigned to UnDef.
1192



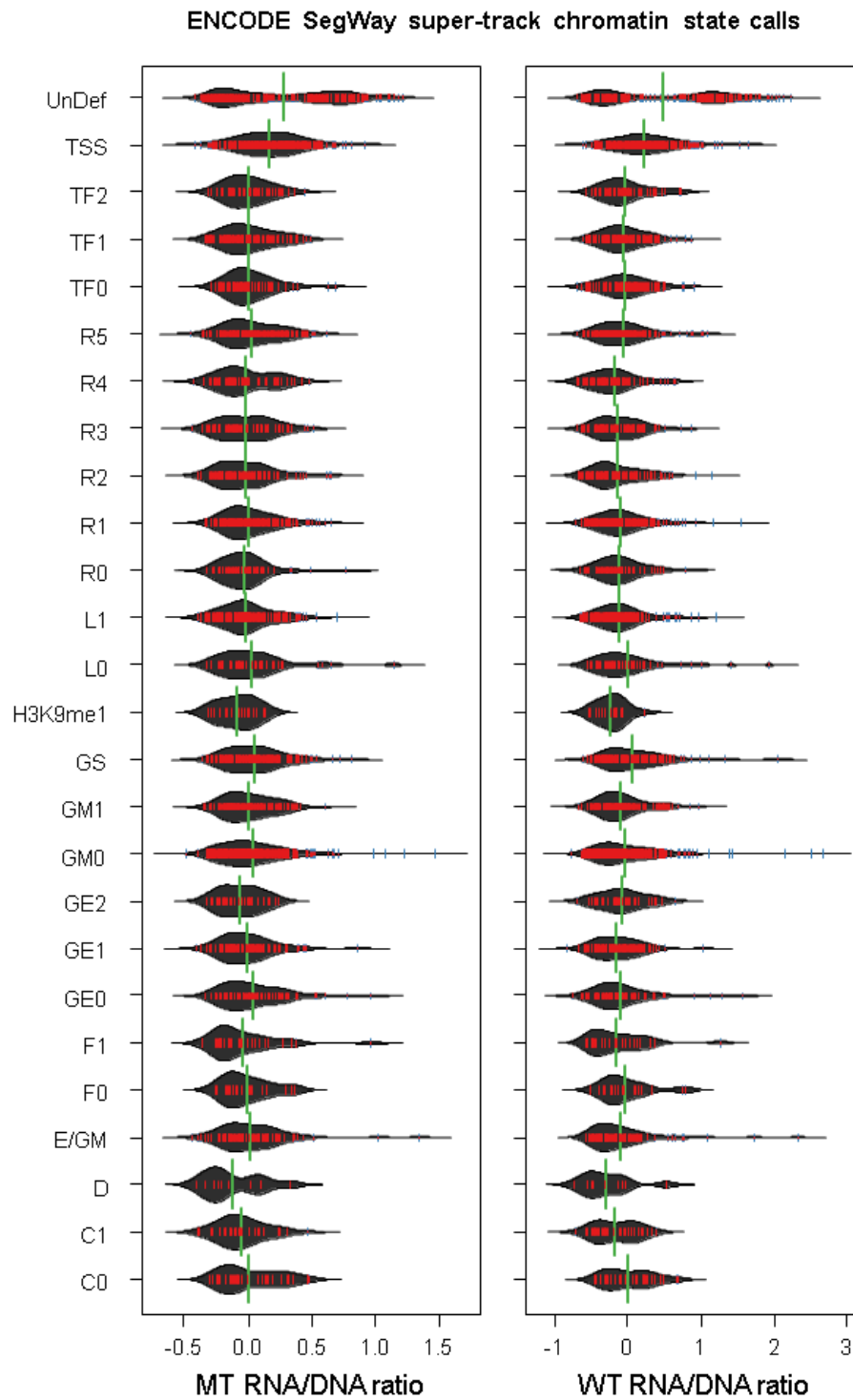
1193

1194 **Supplemental Figure S14.** Distribution of RNA/DNA ratios for MT (top) and WT (bottom) split by HepG2
1195 Open Chromatin states. Chromatin labels were downloaded from the ENCODE project. Inserts not
1196 represented in the available annotations were assigned to UnDef.
1197



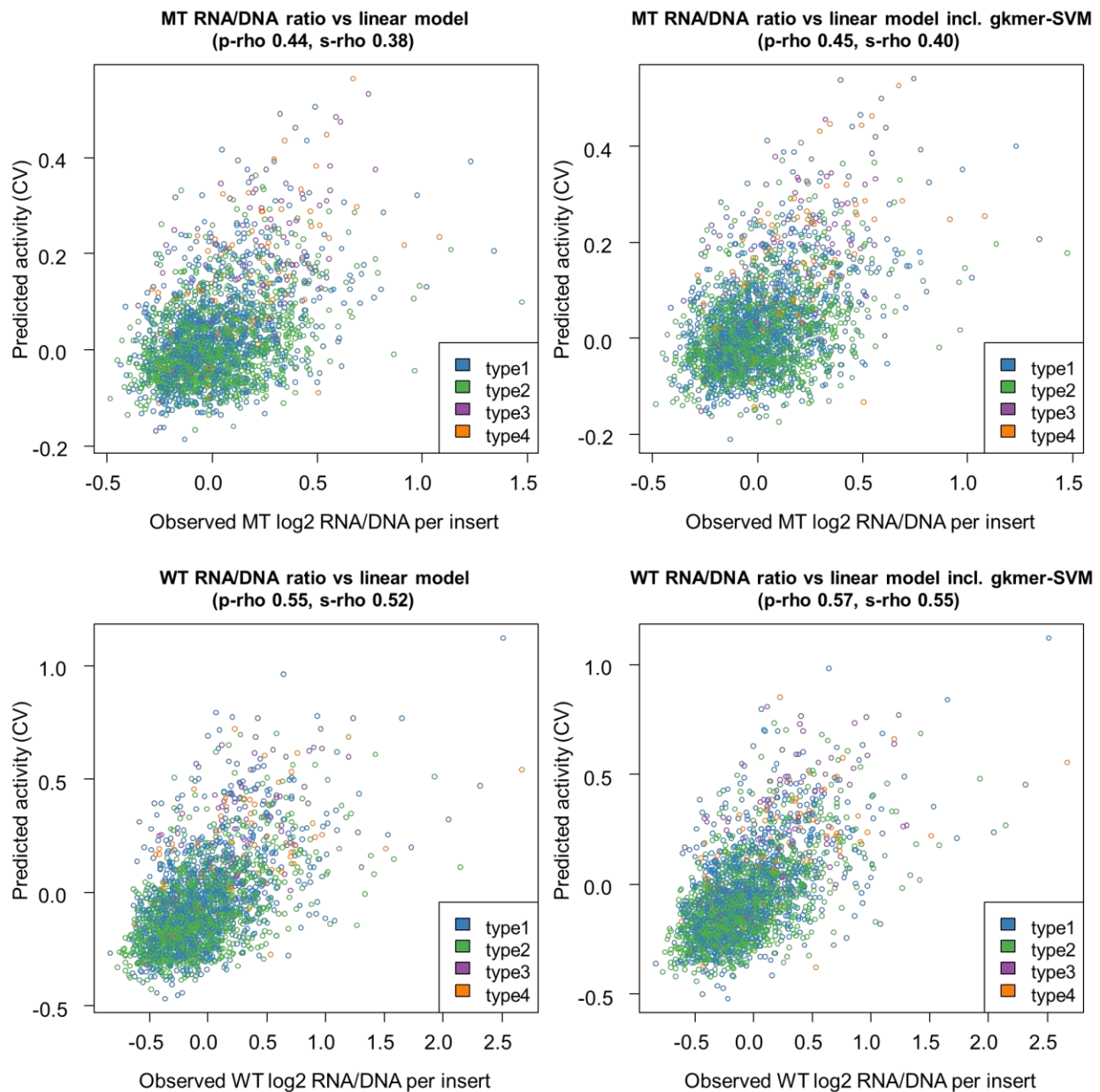
1198
1199

1200 **Supplemental Figure S15.** Distribution of RNA/DNA ratios for the MT (left) and WT (right) experiments
1201 split by SegWay chromatin states of the UCSC supertrack. Inserts not represented in the available
1202 annotations were assigned to UnDef.
1203



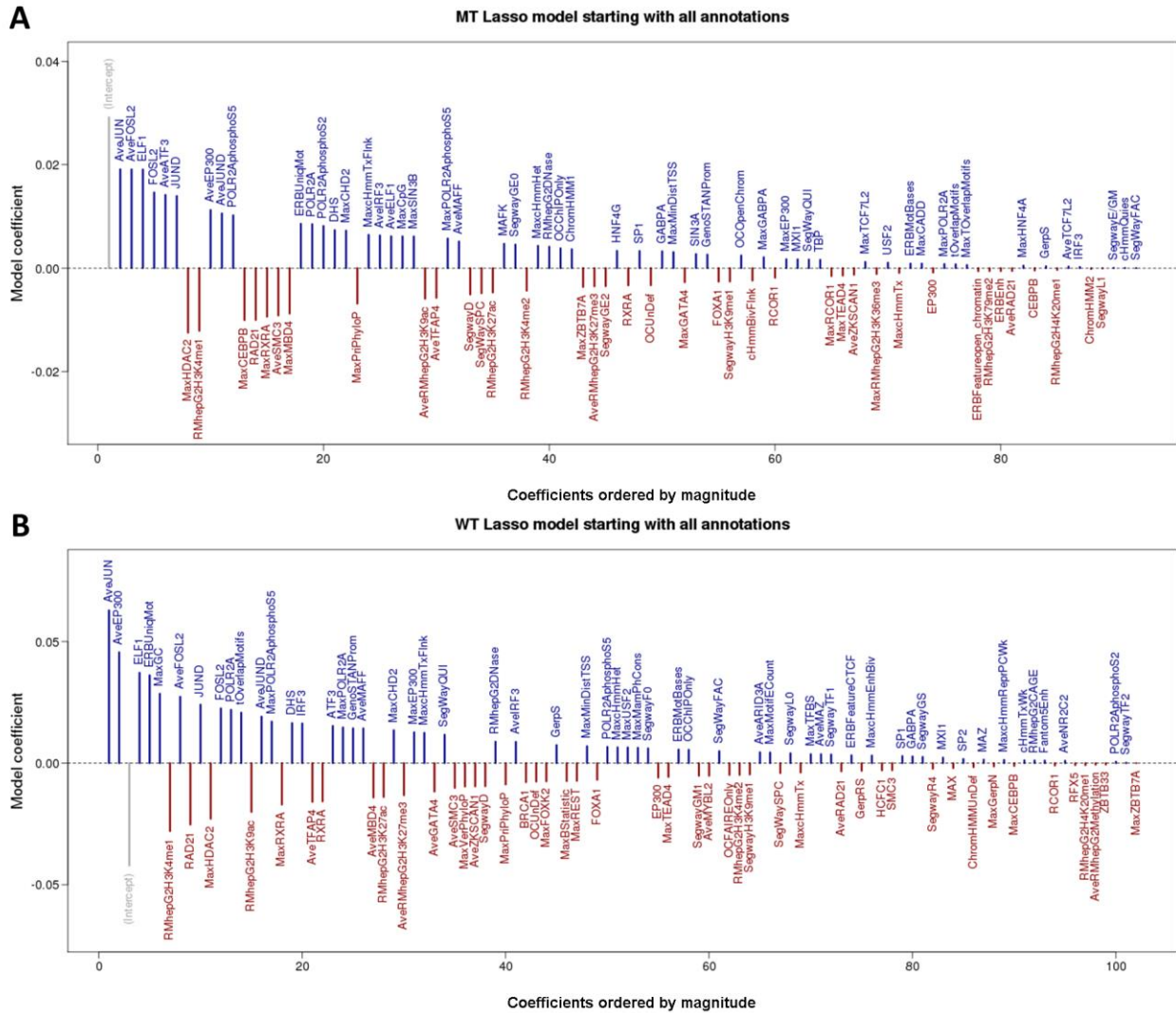
1204

1205 **Supplemental Figure S16.** Correlation (Pearson/p-rho and Spearman/s-rho) of MT and WT RNA/DNA
1206 ratios fit in linear models derived from all genomic annotations with and without HepG2 gkm-SVM scores.
1207 SRES and other control sequences were excluded from this analysis as genomic annotations are mostly
1208 missing for those. WT linear models correlate better with the WT ratios than do MT linear models for the
1209 MT ratios (e.g. Spearman R^2 of 0.2717 WT vs. 0.1459 for MT / Pearson R^2 of 0.3069 for WT vs. 0.1931
1210 for MT). Gapped-kmer SVM scores further improve R^2 values especially for WT ratios (Spearman R^2
1211 0.2975 for WT vs. 0.1581 for MT / Pearson R^2 0.3295 for WT vs. 0.2062 for MT).
1212



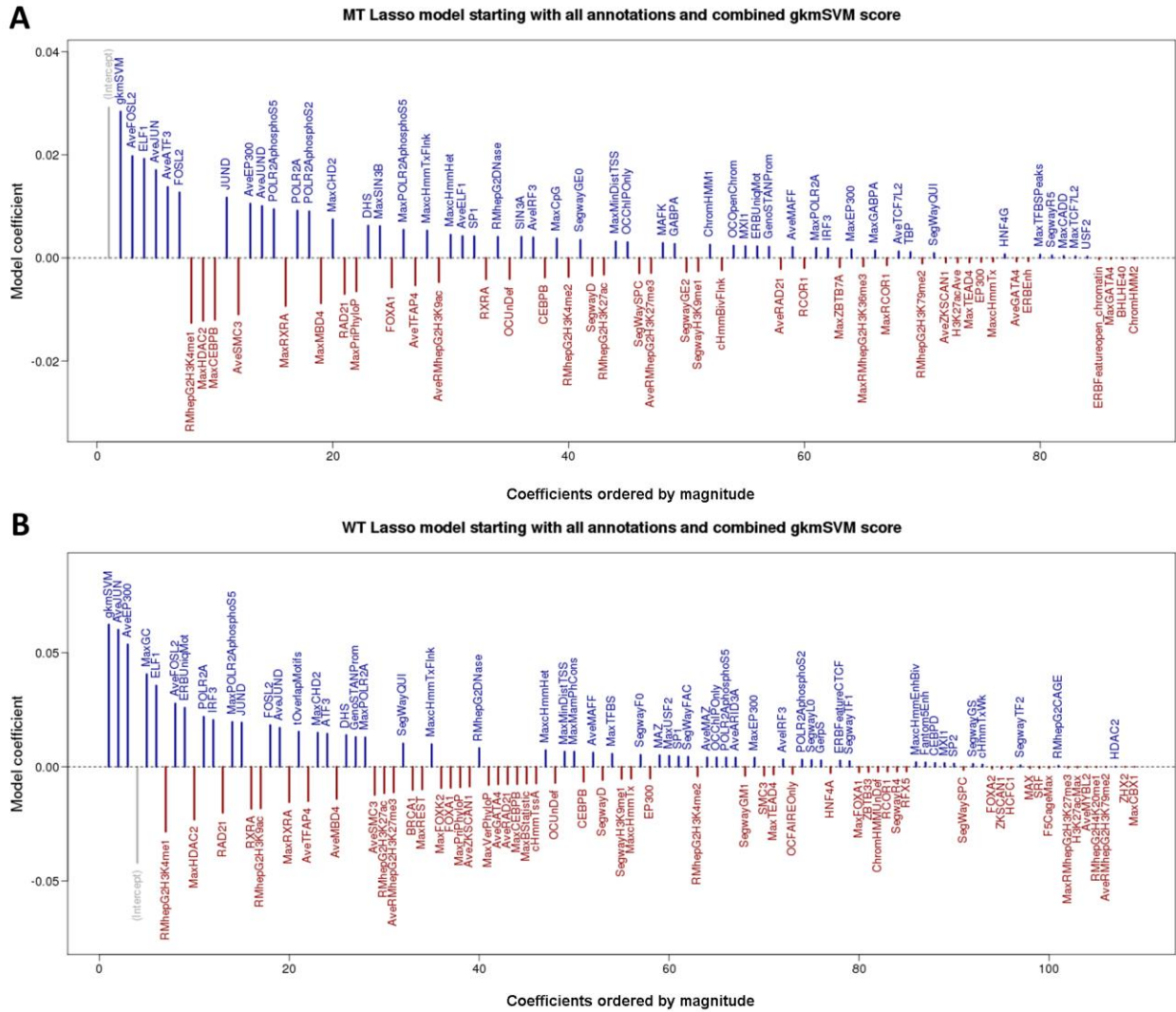
1213

1214 **Supplemental Figure S17.** Coefficients for lasso models to predict RNA/DNA for MT (panel A) and WT
 1215 (panel B) using genomic annotations, without including gkm-SVM score. The R glmnet package was used
 1216 to fit the models, and the tuning parameter for each model was selected via 10-fold cross-validation.
 1217 Categorical features were coded as K-1 binary columns, where K is the number of levels of the categorical
 1218 feature. We excluded ZNF274 and EZH2 annotations from the model, because none of the inserts
 1219 overlapped with these ChIP-seq tracks. All annotation features were scaled and centered before fitting the
 1220 lasso model.
 1221



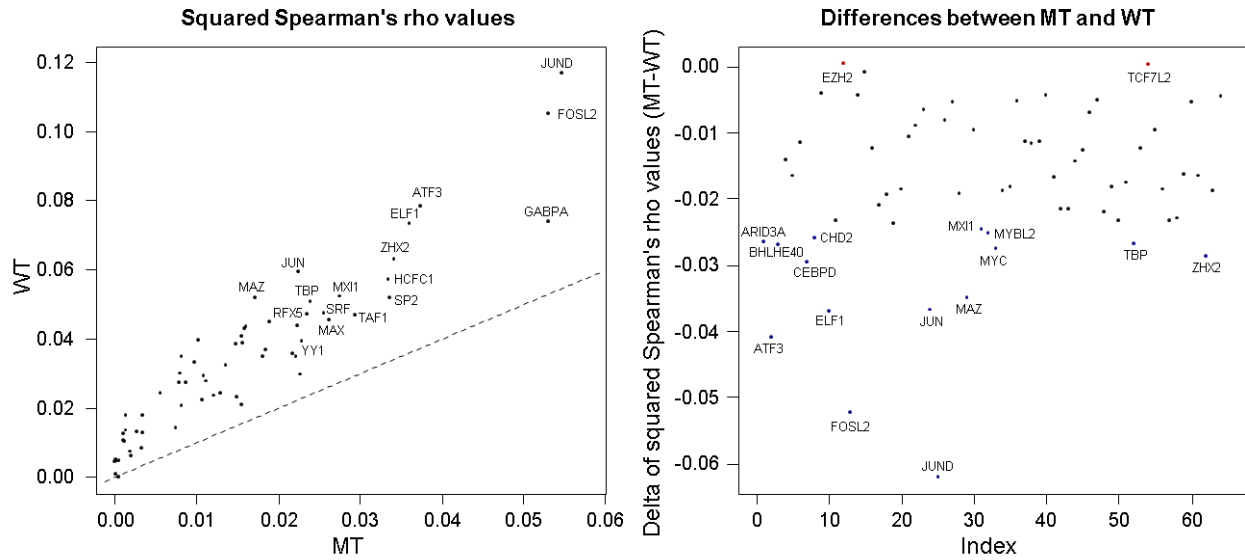
1222
 1223

1224 **Supplemental Figure S18.** Coefficients for lasso models to predict RNA/DNA for MT (panel A) and WT
 1225 (panel B) using genomic annotations and the combined HepG2 gkm-SVM score as predictors. Additional
 1226 details are as in Figure S17. In the resulting models, the sequence based gkm-SVM score is assigned the
 1227 highest model coefficient.



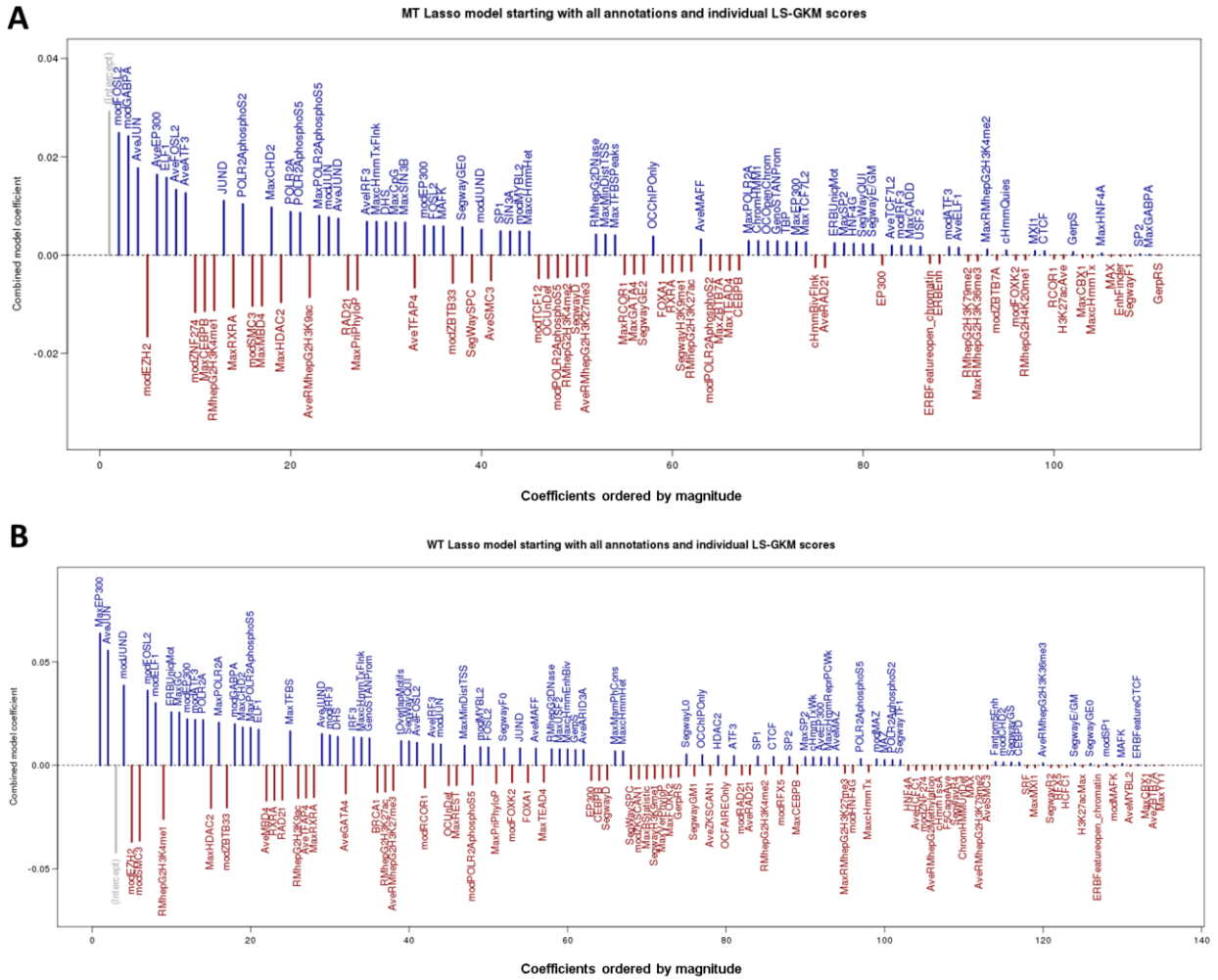
1228
 1229
 1230

1231 **Supplemental Figure S19.** Spearman correlation coefficients (R^2) of measured non-control insert activity
1232 with 64 LS-GKM models trained from HepG2 ChIP-Seq data. WT RNA/DNA ratios correlate better with
1233 annotations than the respective MT values. The left panel highlights the top correlated annotations for WT
1234 and MT ratios. The right panel highlights annotations with the largest difference in R^2 values between MT
1235 and WT.



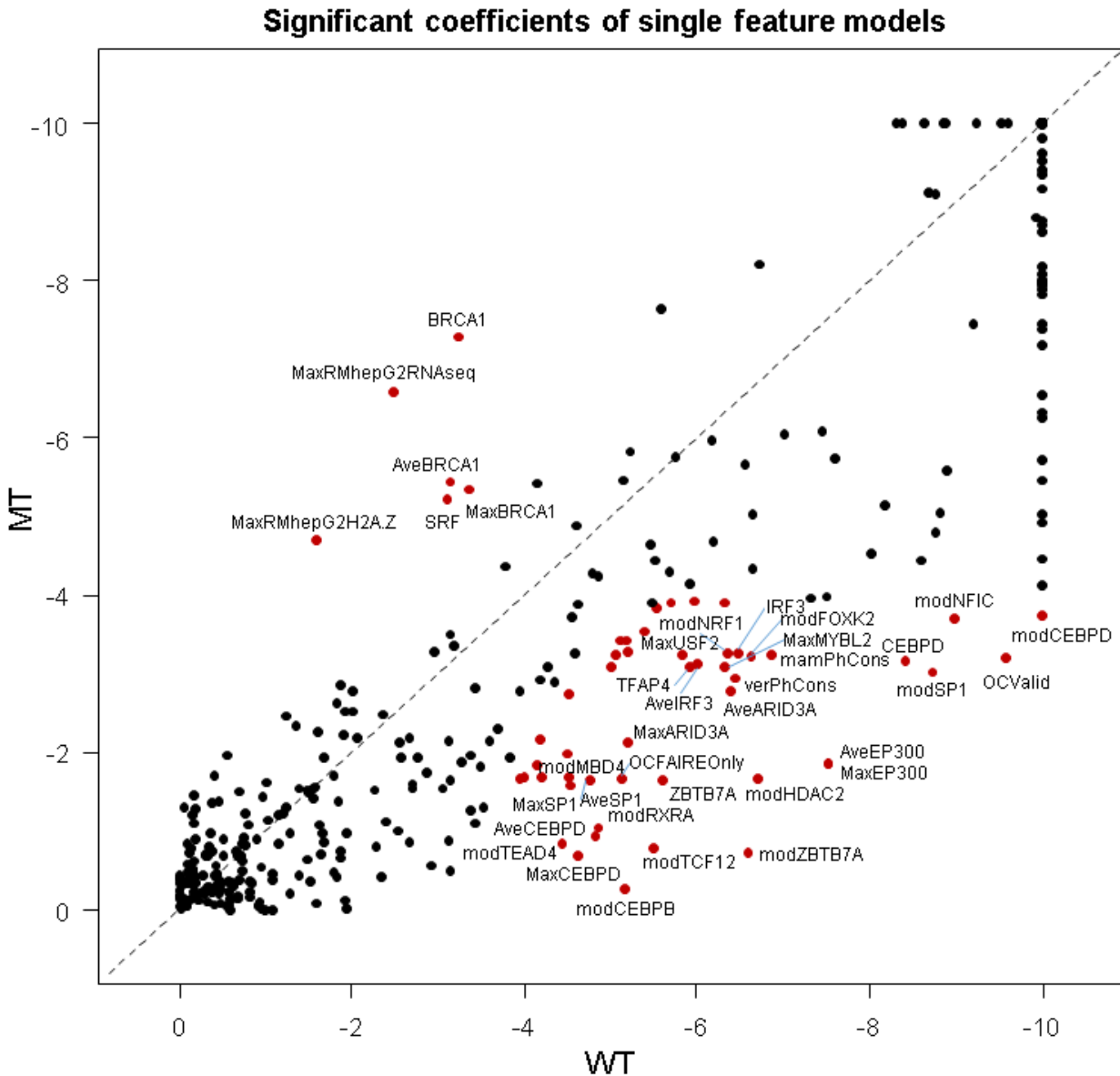
1236
1237
1238

1239 **Supplemental Figure S20.** Coefficients for a lasso model that uses genomic annotations as well as LS-
1240 GKM scores to predict RNA/DNA ratios. Models were fit as described in the methods and summarized in
1241 Fig. S17. In addition, the 64 HepG2 ChIP-seq derived LS-GKM model scores were included as predictors
1242 (Mod-prefix).



1243
1244

1245 **Supplemental Figure S21.** The results of 440 single-feature linear models predicting log2 RNA/DNA
1246 ratios using a single genomic annotation. The two-tailed p-value corresponding to the t-ratio based on a
1247 Student-t distribution (Table S7) is plotted for the inclusion of each coefficient in a single coefficient plus
1248 intercept linear model for predicting log2 RNA/DNA ratios for MT and WT experiments. For plotting
1249 purposes, p-values smaller than 10^{-10} were set to this threshold and p-values were log10 transformed.
1250 Highlighted in red are coefficients passing a p-value threshold of 0.05 (0.00012 after Bonferroni correction)
1251 in MT or WT while failing it in the other experiments. We also require a minimum p-value difference
1252 (0.0025) between MT and WT to account for stochasticity in the p-value estimates.
1253



1256
1257
1258
1259
1260
1261
1262
1263
1264
1265
1266

Supplementary tables

Supplemental Table S1. Active sequences by design category. The four classes of putative enhancer elements are: Regions of FOXA1, FOXA2 or HNF4A binding that overlap H3K27ac and EP300 calls as well as at least one of three chromatin remodeling factors RAD21, CHD2 or SMC3 (type 1); Regions like in 1 but with no remodeling factor overlapping (type 2); EP300 peak regions overlapping H3K27ac as well as at least one of chromatin remodeling factor, but without peaks in FOXA1, FOXA2 or HNF4A (type 3); Regions like in 3 but with no remodeling factor overlapping (type 4). Active sequences are defined as being above the 90th percentile of RNA/DNA ratios observed for the negative synthetic controls (SRES). Promoters are defined based on a value less or equal to 1000 in the minDistTSS annotation field.

	Design categories				Pro moter	Other	Excluding putative promoters			
	1	2	3	4			1	2	3	4
Active	373	328	62	47	134	676	332	297	23	24
WT	36.2%	31.8%	68.9%	54.0%	64.4%	33.3%	34.7%	30.3%	56.1%	46.2%
Remainder	656	702	28	40	74	1352	624	682	18	28
Active	270	266	51	39	112	514	238	243	18	15
MT	26.2%	25.8%	56.7%	44.8%	53.8%	25.3%	24.9%	24.8%	43.9%	28.8%
Remainder	759	764	39	48	96	1514	718	736	23	37

1267
1268

1269 **Supplemental Table S2.** Performance of pooled sequence models measured as Spearman R^2 with
1270 RNA/DNA ratios of MT and WT experiments. The column "gkmSVM" refers to the sequence-based model
1271 trained from combining individual ChIP-seq binding factors described by M Ghandi & D Lee et al. (Ghandi
1272 et al. 2014). JUND is the LS-GKM model trained only on JUND ChIP-seq peaks. This is the model showing
1273 the highest correlation with MT and WT RNA/DNA ratios. "Top 5", "Top 10" and "All Lasso" LS-GKM
1274 models were trained on the pooled ChIP-seq data underlying the individual TF LS-GKM models selected
1275 by linear Lasso regression. The "Top 5" factors are FOSL2, JUND, GABPA, EZH2, SMC3 (MT) as well
1276 as JUND, FOSL2, EZH2, JUN, and SMC3 (WT). The "Top 10" additionally includes MYBL2, FOXA2,
1277 CEBPB, JUN, ATF3 (MT) as well as ELF1, MYBL2, ATF3 GABPA, ZBTB33 (WT). The full set of factors
1278 used for "All Lasso" (in alphabetical order) is ATF3, BRCA1, CEBPB, ELF1, EZH2, FOSL2, FOXA1,
1279 FOXA2, FOXK2, GABPA, HCFC1, HNF4G, IRF3, JUN, JUND, MAFF, MYBL2, NR2C2, POLR2A,
1280 POLR2AphosphoS2, POLR2AphosphoS5, RAD21, RCOR1, RFX5, SIN3B, SMC3, SP2, SRF, TCF12,
1281 TEAD4, TFAP4, ZBTB33, ZBTB7A, ZNF274 (MT) and ARID3A, ATF3, CEBPB, CEBPD, CHD2, ELF1,
1282 EZH2, FOSL2, FOXA1, FOXA2, FOXK2, GABPA, HNF4G, IRF3, JUN, JUND, MAFK, MAZ, MYBL2,
1283 MYC, NR2C2, POLR2A, POLR2AphosphoS2, POLR2AphosphoS5, RAD21, RCOR1, RFX5, SIN3A,
1284 SMC3, SP2, SRF, TAF1, TBP, TEAD4, TFAP4, ZBTB33, ZBTB7A, ZKSCAN1 (WT).

1285

Experiment	gkmSVM	JUND	Top 5	Top 10	All Lasso
WT	7.56%	11.68%	13.76%	15.18%	9.39%
MT	4.09%	5.47%	8.16%	5.04%	5.05%

1286

1287 **Supplemental Table S3.** Primer sequences used for cloning, barcode amplification and sequencing.
1288

Name	Sequence
<i>Cloning adaptors:</i>	
pLSmP-AG-f	GGCCCGCTCTAGACCTAGGACCGGATCAACT
pLSmP-AG-r	GTCCCTCGACGAATTGTCGGTTCACGCAATG
<i>Enhancer/Barcode association primers:</i>	
pLSmP-ass-F	AATGATACGGCGACCACCGAGATCTACACCAGCCTGCATTT CTGCCAGGG
pLSmP-ass-R-i#	CAAGCAGAAGACGGCATAACGAGAT*10bpIndx*CACGAAGTT ATTAGGTCCCTCGAC
pLSmP-AG-seqR1	CAGGGCCCGCTCTAGACCTAGGACCGGATCAACT
pLSmP-AG-seqIndx	ACCGACAATTCGTCGAGGGACCTAATAACTTCG
pLSmP-AG-seqR2	TAGGTCCCTCGACGAATTGTCGGTTCACGCAATG
<i>qPCR primers used for DNA titration:</i>	
Genomic.F	TCCTCCGGAGTTATTCTTGGCA
Genomic.R	CCCCCATCTGATCTGTTTCAC
WPRE.F	TACGCTGCTTTAATGCCTTTG
WPRE.R	GGGCCACAACCTCATAAAG
2-LTR.F	GAGTCCTGCGTCGAGAGAGC
2-LTR.R	AACTAGGGAACCCACTGCTTAAG
<i>Barcode readout primers (pLSmP-ass-R-i# and pLSmP-AG-seqIndx are also reused for this):</i>	
BARCODE_lentiF_v4.1	AATGATACGGCGACCACCGAGATCTACACCTCGGCATGGAC GAGCTGTACAAGTAG
BARCODE-SEQ-R1-V4	CTCGGCATGGACGAGCTGTACAAGTAGGAATTC
BARCODE-SEQ-R2-V4	AGGTCCCTCGACGAATTGTCGGTTCACGCAATG

1289

1290 **Supplemental Table S4.** Proportion of correct oligos for increasing number of observations / number of
1291 sequences going into consensus calling. The major effect is due to insertion/deletion (InDels) events in the
1292 oligos, substitutions have a minor effect and are removed in the consensus calling process as long as the
1293 correct sequence is the most abundant sequence.
1294

Min. sequence count	Oligos without InDels	Correct oligos	All oligos	Percent InDels	Percent correct
0	197752	190093	992513	19.92%	19.15%
5	191537	184487	870677	22.00%	21.19%
10	171625	166177	547199	31.36%	30.37%
15	151109	147218	312297	48.39%	47.14%
20	132546	129953	188893	70.17%	68.80%
25	115851	114238	132995	87.11%	85.90%
30	101675	100728	106507	95.46%	94.57%
35	89282	88754	90699	98.44%	97.86%
40	78521	78224	78976	99.42%	99.05%
45	69247	69081	69434	99.73%	99.49%
50	61229	61144	61317	99.86%	99.72%
55	54256	54212	54300	99.92%	99.84%
60	48228	48205	48253	99.95%	99.90%
65	43094	43076	43110	99.96%	99.92%
70	38746	38730	38755	99.98%	99.94%
75	34866	34853	34873	99.98%	99.94%
80	31434	31426	31437	99.99%	99.97%
85	28429	28423	28431	99.99%	99.97%
90	25801	25795	25803	99.99%	99.97%
95	23502	23497	23503	100.00%	99.97%
100	21422	21418	21423	100.00%	99.98%

1295
1296

1297 **Supplemental Table S5.** Primers used for cloning 2 negative and 2 positive control sequences in Luciferase
1298 Assay experiments.
1299

Name	Sequence
pos1.F	GCTAGCCTCGAGGATATGCAGCTGTCAAGAAAAATGAG
pos1.R	TCTAGTGTCTAAGCTAGGAGGAGGGGCTGATACAG
pos2.F	GCTAGCCTCGAGGATCGCTTCCGGCCACTTGGCC
pos2.R	TCTAGTGTCTAAGCTCTCCCCGCTGGCTTCCTAC
neg1.F	GCTAGCCTCGAGGATCAGGATGCCCTGGCCAGTG
neg1.R	TCTAGTGTCTAAGCTCCTATGGATTCAATACTGACTG
neg2.F	GCTAGCCTCGAGGATGGGCCAGCGCCTTAAATGAC
neg2.R	TCTAGTGTCTAAGCTGAGTCAAGGACATAAGCATGC

1300
1301

1302 **Supplemental Table S6.** LS-GKM training on 225,327 HepG2 combined ChIP-seq peaks (positive set)
1303 and a positive validation set of 10,000 left-out sequences. Training was performed using LS-GKM defaults
1304 (DEF, -T 4 -e 0.01) and parameters matching gkm-SVM (GKM, -l 10 -k 6 -d 3 -t 2 -T 4 -e 0.01). Negative
1305 sequences were either created by permuting the positive sequence while maintaining dimer-content or used
1306 the negative sequence set provided with gkm-SVM.
1307

Name	TP	FN	TN	FP	ACC	PPV	TPR	SPC
GKM:pos-neg	8217	1783	7557	2443	78.9%	77.1%	82.2%	75.6%
GKM:pos-shuf	9055	945	6443	3557	77.5%	71.8%	90.6%	64.4%
DEF:pos-neg	8631	1369	8033	1967	83.3%	81.4%	86.3%	80.3%
DEF:pos-shuf	9190	810	7042	2958	81.2%	75.7%	91.9%	70.4%

1308
1309

1310 **Supplemental Table S7.** Model coefficients determined for 440 one feature models (single coefficient plus
 1311 offset) for predicting log₂ RNA/DNA ratios for MT and WT experiments. Categorical features were
 1312 included as n-1 binary columns, where n is the number of levels of the categorical feature. ZNF274 and
 1313 EZH2 annotations (but not the sequence based models modZNF274 and modEZH2) as none of the inserts
 1314 overlapped with these ChIP-seq tracks. All annotation features and the output variable (MT/WT log₂
 1315 RNA/DNA ratios) were scaled and centered before fitting the models to allow interpretation of coefficient
 1316 values. The table provides the coefficient estimate (Coeff.), standard error (Std. Err.), t-values, and Pr(>|t|),
 1317 which is the two-tailed p-value corresponding to the t-ratio based on a Student-t distribution. These values
 1318 are returned by the R glm.summary routine for each model.
 1319

Annotation feature	MT				WT			
	Coeff.	Std. Err.	t value	Pr(> t)	Coeff.	Std. Err.	t value	Pr(> t)
SimpleRepeat	0.016	0.021	0.750	4.535E-01	0.000	0.021	-0.020	9.837E-01
DHS	0.073	0.021	3.463	5.439E-04	0.089	0.021	4.215	2.601E-05
GenoSTANProm	0.051	0.021	2.428	1.524E-02	0.076	0.021	3.601	3.234E-04
ERBProm	0.163	0.021	7.801	9.325E-15	0.172	0.021	8.257	2.532E-16
Fantom5Enh	0.049	0.021	2.307	2.112E-02	0.085	0.021	4.012	6.230E-05
GenoSTANEnh	0.013	0.021	0.635	5.258E-01	0.015	0.021	0.727	4.676E-01
EnhFinder	-0.014	0.021	-0.657	5.111E-01	0.019	0.021	0.902	3.671E-01
ERBEnh	-0.044	0.021	-2.072	3.838E-02	-0.046	0.021	-2.198	2.808E-02
ERBUniqMot	0.270	0.020	13.241	1.392E-38	0.360	0.020	18.263	1.456E-69
ERBMotBases	0.219	0.021	10.634	8.481E-26	0.289	0.020	14.285	2.303E-44
H3K27ac	0.003	0.021	0.146	8.840E-01	0.018	0.021	0.861	3.894E-01
H3K27acAve	0.034	0.021	1.611	1.072E-01	0.049	0.021	2.308	2.110E-02
H3K27acMax	0.053	0.021	2.523	1.170E-02	0.063	0.021	3.006	2.675E-03
F5Cage	0.169	0.021	8.083	1.023E-15	0.136	0.021	6.491	1.046E-10
F5CageAve	0.044	0.021	2.100	3.584E-02	0.034	0.021	1.627	1.039E-01
F5CageMax	0.028	0.021	1.314	1.891E-01	0.006	0.021	0.301	7.635E-01
TFcount	0.220	0.021	10.651	7.111E-26	0.268	0.020	13.144	4.599E-38
ARID3A	0.049	0.021	2.299	2.158E-02	0.082	0.021	3.873	1.108E-04
ATF3	0.182	0.021	8.766	3.583E-18	0.219	0.021	10.625	9.311E-26
BHLHE40	0.087	0.021	4.135	3.682E-05	0.126	0.021	5.983	2.550E-09
BRCA1	0.115	0.021	5.462	5.239E-08	0.073	0.021	3.451	5.692E-04
CBX1	0.154	0.021	7.361	2.545E-13	0.158	0.021	7.557	5.996E-14
CEBPB	-0.037	0.021	-1.752	7.987E-02	0.009	0.021	0.429	6.681E-01
CEBPD	0.072	0.021	3.397	6.941E-04	0.124	0.021	5.917	3.788E-09
CHD2	0.152	0.021	7.279	4.651E-13	0.180	0.021	8.625	1.197E-17
CTCF	0.031	0.021	1.455	1.458E-01	0.038	0.021	1.809	7.051E-02
ELF1	0.202	0.021	9.734	5.866E-22	0.216	0.021	10.470	4.458E-25
EP300	-0.006	0.021	-0.278	7.809E-01	0.002	0.021	0.102	9.187E-01
FOSL2	0.241	0.021	11.730	7.043E-31	0.283	0.020	13.958	1.636E-42
FOXA1	-0.133	0.021	-6.364	2.373E-10	-0.147	0.021	-7.024	2.841E-12
FOXA2	-0.104	0.021	-4.925	9.042E-07	-0.112	0.021	-5.347	9.876E-08
FOKK2	0.063	0.021	2.969	3.017E-03	0.053	0.021	2.524	1.169E-02
GABPA	0.215	0.021	10.423	7.134E-25	0.203	0.021	9.781	3.749E-22
GATA4	0.087	0.021	4.132	3.723E-05	0.099	0.021	4.681	3.030E-06
HCFC1	0.182	0.021	8.765	3.631E-18	0.147	0.021	7.044	2.471E-12
HDAC2	-0.008	0.021	-0.362	7.171E-01	0.027	0.021	1.256	2.093E-01
HNF4A	-0.016	0.021	-0.779	4.363E-01	-0.021	0.021	-0.977	3.288E-01
HNF4G	0.008	0.021	0.383	7.020E-01	0.002	0.021	0.115	9.085E-01
IRF3	0.073	0.021	3.461	5.484E-04	0.108	0.021	5.123	3.256E-07
JUN	0.173	0.021	8.303	1.736E-16	0.268	0.020	13.138	4.915E-38
JUND	0.229	0.021	11.132	4.701E-28	0.277	0.020	13.635	1.008E-40

Annotation feature	MT				WT			
	Coeff.	Std. Err.	t value	Pr(> t)	Coeff.	Std. Err.	t value	Pr(> t)
MAFF	0.092	0.021	4.371	1.292E-05	0.089	0.021	4.227	2.468E-05
MAFK	0.098	0.021	4.654	3.439E-06	0.095	0.021	4.501	7.125E-06
MAX	0.116	0.021	5.524	3.704E-08	0.130	0.021	6.213	6.195E-10
MAZ	0.143	0.021	6.827	1.116E-11	0.203	0.021	9.780	3.793E-22
MBD4	0.000	0.021	0.014	9.891E-01	0.037	0.021	1.735	8.289E-02
MXI1	0.185	0.021	8.912	1.015E-18	0.207	0.021	10.016	3.972E-23
MYBL2	0.082	0.021	3.889	1.035E-04	0.117	0.021	5.556	3.080E-08
MYC	0.090	0.021	4.268	2.056E-05	0.105	0.021	4.994	6.381E-07
NFIC	0.052	0.021	2.470	1.360E-02	0.073	0.021	3.471	5.285E-04
NR2C2	0.067	0.021	3.173	1.528E-03	0.075	0.021	3.566	3.699E-04
NRF1	0.079	0.021	3.742	1.873E-04	0.088	0.021	4.197	2.816E-05
POLR2A	0.200	0.021	9.666	1.114E-21	0.228	0.021	11.075	8.672E-28
POLR2A phosphoS2	0.186	0.021	8.950	7.268E-19	0.165	0.021	7.885	4.875E-15
POLR2 AphosphoS5	0.198	0.021	9.557	3.075E-21	0.194	0.021	9.344	2.172E-20
RAD21	-0.027	0.021	-1.260	2.079E-01	-0.015	0.021	-0.697	4.860E-01
RCOR1	0.071	0.021	3.347	8.315E-04	0.093	0.021	4.431	9.850E-06
REST	0.160	0.021	7.666	2.620E-14	0.169	0.021	8.081	1.043E-15
RFX5	0.122	0.021	5.830	6.335E-09	0.110	0.021	5.226	1.894E-07
RXRA	-0.025	0.021	-1.169	2.426E-01	-0.009	0.021	-0.411	6.813E-01
SIN3A	0.163	0.021	7.807	8.916E-15	0.154	0.021	7.363	2.522E-13
SIN3B	0.199	0.021	9.608	1.913E-21	0.158	0.021	7.565	5.613E-14
SMC3	-0.007	0.021	-0.336	7.369E-01	0.024	0.021	1.156	2.476E-01
SP1	0.026	0.021	1.224	2.212E-01	0.027	0.021	1.300	1.937E-01
SP2	0.073	0.021	3.468	5.348E-04	0.069	0.021	3.271	1.088E-03
SRF	0.096	0.021	4.537	6.019E-06	0.071	0.021	3.367	7.725E-04
TAF1	0.188	0.021	9.051	3.005E-19	0.196	0.021	9.439	9.101E-21
TBP	0.174	0.021	8.349	1.191E-16	0.171	0.021	8.221	3.378E-16
TCF12	0.041	0.021	1.954	5.079E-02	0.077	0.021	3.628	2.915E-04
TCF7L2	0.015	0.021	0.729	4.661E-01	0.009	0.021	0.445	6.565E-01
TEAD4	-0.003	0.021	-0.152	8.796E-01	0.023	0.021	1.099	2.721E-01
TFAP4	0.071	0.021	3.342	8.467E-04	0.102	0.021	4.869	1.198E-06
USF1	0.071	0.021	3.346	8.345E-04	0.085	0.021	4.051	5.274E-05
USF2	0.081	0.021	3.844	1.242E-04	0.098	0.021	4.664	3.281E-06
YY1	0.149	0.021	7.118	1.468E-12	0.170	0.021	8.146	6.213E-16
ZBTB33	0.048	0.021	2.256	2.416E-02	0.035	0.021	1.660	9.698E-02
ZBTB7A	0.048	0.021	2.277	2.290E-02	0.099	0.021	4.725	2.440E-06
ZHX2	0.144	0.021	6.868	8.403E-12	0.145	0.021	6.946	4.919E-12
ZKSCAN1	0.060	0.021	2.835	4.618E-03	0.043	0.021	2.012	4.433E-02
AveARID3A	0.066	0.021	3.142	1.703E-03	0.107	0.021	5.085	3.988E-07
AveATF3	0.199	0.021	9.575	2.602E-21	0.178	0.021	8.554	2.165E-17
AveBHLHE40	0.080	0.021	3.807	1.445E-04	0.099	0.021	4.690	2.895E-06
AveBRCA1	0.098	0.021	4.641	3.667E-06	0.072	0.021	3.393	7.038E-04
AveCBX1	0.129	0.021	6.172	8.009E-10	0.127	0.021	6.046	1.732E-09
AveCEBPB	-0.043	0.021	-2.033	4.222E-02	0.020	0.021	0.943	3.458E-01
AveCEBPD	0.033	0.021	1.578	1.147E-01	0.091	0.021	4.343	1.469E-05
AveCHD2	0.127	0.021	6.062	1.571E-09	0.136	0.021	6.474	1.170E-10
AveCTCF	0.007	0.021	0.343	7.319E-01	0.024	0.021	1.143	2.533E-01
AveELF1	0.231	0.021	11.235	1.572E-28	0.209	0.021	10.097	1.806E-23
AveEP300	0.052	0.021	2.462	1.391E-02	0.117	0.021	5.560	3.018E-08
AveFOSL2	0.276	0.020	13.557	2.690E-40	0.358	0.020	18.117	1.487E-68
AveFOXA1	-0.026	0.021	-1.253	2.104E-01	-0.022	0.021	-1.037	3.001E-01
AveFOXA2	-0.038	0.021	-1.802	7.164E-02	-0.035	0.021	-1.677	9.372E-02

Annotation feature	MT				WT			
	Coeff.	Std. Err.	t value	Pr(> t)	Coeff.	Std. Err.	t value	Pr(> t)
AveFOXK2	0.063	0.021	2.963	3.078E-03	0.055	0.021	2.586	9.765E-03
AveGABPA	0.204	0.021	9.868	1.642E-22	0.167	0.021	7.998	2.002E-15
AveGATA4	0.027	0.021	1.279	2.011E-01	0.048	0.021	2.291	2.204E-02
AveHCFC1	0.172	0.021	8.275	2.186E-16	0.123	0.021	5.874	4.891E-09
AveHDAC2	-0.027	0.021	-1.286	1.986E-01	0.015	0.021	0.717	4.732E-01
AveHNF4A	0.034	0.021	1.624	1.045E-01	0.053	0.021	2.530	1.148E-02
AveHNF4G	0.019	0.021	0.910	3.632E-01	0.043	0.021	2.044	4.105E-02
AveIRF3	0.071	0.021	3.372	7.601E-04	0.103	0.021	4.907	9.908E-07
AveJUN	0.224	0.021	10.872	7.276E-27	0.328	0.020	16.422	2.656E-57
AveJUND	0.300	0.020	14.843	1.329E-47	0.398	0.019	20.529	6.186E-86
AveMAFF	0.086	0.021	4.080	4.653E-05	0.109	0.021	5.197	2.213E-07
AveMAFK	0.057	0.021	2.679	7.441E-03	0.077	0.021	3.667	2.511E-04
AveMAX	0.143	0.021	6.819	1.175E-11	0.154	0.021	7.371	2.370E-13
AveMAZ	0.159	0.021	7.612	3.942E-14	0.204	0.021	9.865	1.692E-22
AveMBD4	0.003	0.021	0.161	8.722E-01	0.033	0.021	1.550	1.213E-01
AveMXI1	0.206	0.021	9.957	7.020E-23	0.211	0.021	10.200	6.569E-24
AveMYBL2	0.073	0.021	3.475	5.204E-04	0.095	0.021	4.534	6.084E-06
AveMYC	0.101	0.021	4.785	1.821E-06	0.117	0.021	5.590	2.543E-08
AveNFIC	0.047	0.021	2.237	2.537E-02	0.065	0.021	3.098	1.975E-03
AveNR2C2	0.074	0.021	3.522	4.377E-04	0.072	0.021	3.407	6.688E-04
AveNRF1	0.068	0.021	3.243	1.202E-03	0.084	0.021	4.001	6.506E-05
AvePOLR2A	0.232	0.021	11.291	8.628E-29	0.252	0.020	12.319	8.705E-34
AvePOLR2A phosphoS2	0.175	0.021	8.399	7.864E-17	0.140	0.021	6.692	2.768E-11
AvePOLR2 A phosphoS5	0.187	0.021	8.982	5.498E-19	0.183	0.021	8.814	2.385E-18
AveRAD21	-0.020	0.021	-0.962	3.362E-01	-0.006	0.021	-0.281	7.785E-01
AveRCOR1	0.081	0.021	3.835	1.293E-04	0.100	0.021	4.764	2.017E-06
AveREST	0.153	0.021	7.313	3.631E-13	0.128	0.021	6.088	1.338E-09
AveRFX5	0.089	0.021	4.246	2.266E-05	0.098	0.021	4.654	3.445E-06
AveRXRA	-0.016	0.021	-0.744	4.571E-01	0.021	0.021	1.014	3.108E-01
AveSIN3A	0.175	0.021	8.412	7.097E-17	0.161	0.021	7.713	1.836E-14
AveSIN3B	0.184	0.021	8.840	1.900E-18	0.126	0.021	6.003	2.259E-09
AveSMC3	-0.031	0.021	-1.446	1.484E-01	0.006	0.021	0.272	7.860E-01
AveSP1	0.048	0.021	2.281	2.265E-02	0.091	0.021	4.311	1.694E-05
AveSP2	0.062	0.021	2.941	3.307E-03	0.060	0.021	2.849	4.428E-03
AveSRF	0.068	0.021	3.202	1.382E-03	0.052	0.021	2.478	1.327E-02
AveTAF1	0.195	0.021	9.402	1.276E-20	0.186	0.021	8.931	8.586E-19
AveTBP	0.154	0.021	7.382	2.196E-13	0.153	0.021	7.318	3.483E-13
AveTCF12	0.037	0.021	1.757	7.906E-02	0.075	0.021	3.566	6.699E-04
AveTCF7L2	0.026	0.021	1.245	2.133E-01	0.026	0.021	1.251	2.112E-01
AveTEAD4	-0.011	0.021	-0.512	6.087E-01	0.031	0.021	1.447	1.479E-01
AveTFAP4	0.020	0.021	0.944	3.452E-01	0.037	0.021	1.740	8.208E-02
AveUSF1	0.086	0.021	4.067	4.930E-05	0.100	0.021	4.757	2.088E-06
AveUSF2	0.084	0.021	3.971	7.370E-05	0.102	0.021	4.868	1.206E-06
AveYY1	0.155	0.021	7.411	1.773E-13	0.124	0.021	5.901	4.155E-09
AveZBTB33	0.059	0.021	2.773	5.597E-03	0.047	0.021	2.248	2.470E-02
AveZBTB7A	0.023	0.021	1.087	2.770E-01	0.069	0.021	3.249	1.175E-03
AveZHX2	0.176	0.021	8.436	5.817E-17	0.140	0.021	6.666	3.298E-11
AveZKSCAN1	0.022	0.021	1.024	3.058E-01	0.008	0.021	0.390	6.966E-01
MaxARID3A	0.056	0.021	2.666	7.741E-03	0.095	0.021	4.533	6.130E-06
MaxATF3	0.198	0.021	9.533	3.822E-21	0.176	0.021	8.451	5.132E-17
MaxBHLHE40	0.075	0.021	3.560	3.784E-04	0.095	0.021	4.489	7.529E-06
MaxBRCA1	0.097	0.021	4.593	4.605E-06	0.074	0.021	3.530	4.247E-04

Annotation feature	MT				WT			
	Coeff.	Std. Err.	t value	Pr(> t)	Coeff.	Std. Err.	t value	Pr(> t)
MaxCBX1	0.130	0.021	6.180	7.581E-10	0.126	0.021	6.017	2.064E-09
MaxCEBPB	-0.043	0.021	-2.017	4.382E-02	0.018	0.021	0.828	4.079E-01
MaxCEBPD	0.026	0.021	1.240	2.151E-01	0.089	0.021	4.234	2.384E-05
MaxCHD2	0.131	0.021	6.263	4.520E-10	0.140	0.021	6.677	3.057E-11
MaxCTCF	0.009	0.021	0.431	6.667E-01	0.027	0.021	1.262	2.071E-01
MaxELF1	0.227	0.021	11.018	1.573E-27	0.210	0.021	10.129	1.320E-23
MaxEP300	0.052	0.021	2.464	1.380E-02	0.117	0.021	5.562	2.992E-08
MaxFOSL2	0.266	0.020	13.055	1.365E-37	0.345	0.020	17.366	1.828E-63
MaxFOXA1	-0.031	0.021	-1.459	1.447E-01	-0.028	0.021	-1.325	1.853E-01
MaxFOXA2	-0.046	0.021	-2.173	2.986E-02	-0.045	0.021	-2.132	3.313E-02
MaxFOKK2	0.058	0.021	2.751	5.994E-03	0.053	0.021	2.504	1.236E-02
MaxGABPA	0.207	0.021	10.023	3.707E-23	0.173	0.021	8.300	1.772E-16
MaxGATA4	0.026	0.021	1.220	2.226E-01	0.052	0.021	2.471	1.353E-02
MaxHCFC1	0.177	0.021	8.511	3.119E-17	0.131	0.021	6.227	5.670E-10
MaxHDAC2	-0.032	0.021	-1.523	1.279E-01	0.010	0.021	0.471	6.378E-01
MaxHNF4A	0.031	0.021	1.477	1.397E-01	0.049	0.021	2.318	2.052E-02
MaxHNF4G	0.002	0.021	0.093	9.262E-01	0.033	0.021	1.554	1.204E-01
MaxIRF3	0.057	0.021	2.706	6.866E-03	0.084	0.021	4.000	6.543E-05
MaxJUN	0.203	0.021	9.817	2.661E-22	0.298	0.020	14.760	4.067E-47
MaxJUND	0.280	0.020	13.772	1.776E-41	0.370	0.020	18.839	1.352E-73
MaxMAFF	0.066	0.021	3.121	1.824E-03	0.088	0.021	4.174	3.110E-05
MaxMAFK	0.059	0.021	2.808	5.027E-03	0.079	0.021	3.727	1.985E-04
MaxMAX	0.130	0.021	6.195	6.913E-10	0.156	0.021	7.465	1.188E-13
MaxMAZ	0.151	0.021	7.234	6.420E-13	0.200	0.021	9.628	1.586E-21
MaxMBD4	0.000	0.021	-0.013	9.893E-01	0.035	0.021	1.633	1.027E-01
MaxMXI1	0.195	0.021	9.384	1.512E-20	0.198	0.021	9.528	4.007E-21
MaxMYBL2	0.071	0.021	3.351	8.190E-04	0.106	0.021	5.051	4.761E-07
MaxMYC	0.104	0.021	4.938	8.457E-07	0.116	0.021	5.530	3.566E-08
MaxNFIC	0.046	0.021	2.177	2.960E-02	0.070	0.021	3.327	8.908E-04
MaxNR2C2	0.076	0.021	3.608	3.157E-04	0.072	0.021	3.394	7.008E-04
MaxNRF1	0.067	0.021	3.153	1.639E-03	0.082	0.021	3.874	1.101E-04
MaxPOLR2A	0.233	0.021	11.347	4.704E-29	0.252	0.020	12.330	7.640E-34
MaxPOLR2A phosphoS2	0.170	0.021	8.130	7.010E-16	0.137	0.021	6.524	8.435E-11
MaxPOLR2A phosphoS5	0.232	0.021	11.271	1.073E-28	0.235	0.021	11.431	1.898E-29
MaxRAD21	-0.017	0.021	-0.821	4.118E-01	-0.001	0.021	-0.028	9.774E-01
MaxRCOR1	0.076	0.021	3.626	2.942E-04	0.097	0.021	4.618	4.102E-06
MaxREST	0.153	0.021	7.331	3.177E-13	0.128	0.021	6.083	1.384E-09
MaxRFX5	0.093	0.021	4.439	9.492E-06	0.109	0.021	5.194	2.250E-07
MaxRXRA	-0.020	0.021	-0.928	3.535E-01	0.020	0.021	0.926	3.543E-01
MaxSIN3A	0.183	0.021	8.822	2.217E-18	0.165	0.021	7.917	3.802E-15
MaxSIN3B	0.194	0.021	9.340	2.257E-20	0.133	0.021	6.327	3.001E-10
MaxSMC3	-0.029	0.021	-1.375	1.692E-01	0.007	0.021	0.326	7.447E-01
MaxSP1	0.047	0.021	2.212	2.709E-02	0.088	0.021	4.189	2.914E-05
MaxSP2	0.057	0.021	2.710	6.775E-03	0.056	0.021	2.632	8.557E-03
MaxSRF	0.062	0.021	2.918	3.554E-03	0.040	0.021	1.906	5.683E-02
MaxTAF1	0.199	0.021	9.597	2.114E-21	0.199	0.021	9.590	2.253E-21
MaxTBP	0.150	0.021	7.155	1.126E-12	0.160	0.021	7.683	2.304E-14
MaxTCF12	0.032	0.021	1.498	1.344E-01	0.071	0.021	3.373	7.554E-04
MaxTCF7L2	0.027	0.021	1.263	2.068E-01	0.028	0.021	1.314	1.891E-01
MaxTEAD4	-0.014	0.021	-0.678	4.976E-01	0.025	0.021	1.181	2.376E-01
MaxTFAP4	0.025	0.021	1.183	2.368E-01	0.045	0.021	2.135	3.291E-02
MaxUSF1	0.075	0.021	3.554	3.869E-04	0.095	0.021	4.523	6.420E-06

Annotation feature	MT				WT			
	Coeff.	Std. Err.	t value	Pr(> t)	Coeff.	Std. Err.	t value	Pr(> t)
MaxUSF2	0.073	0.021	3.452	5.660E-04	0.102	0.021	4.827	1.483E-06
MaxYY1	0.157	0.021	7.500	9.144E-14	0.133	0.021	6.355	2.514E-10
MaxZBTB33	0.064	0.021	3.041	2.389E-03	0.052	0.021	2.446	1.452E-02
MaxZBTB7A	0.021	0.021	0.998	3.185E-01	0.072	0.021	3.389	7.130E-04
MaxZHX2	0.182	0.021	8.736	4.655E-18	0.152	0.021	7.270	4.937E-13
MaxZKSCAN1	0.019	0.021	0.875	3.816E-01	0.007	0.021	0.325	7.450E-01
RMhепG2	0.146	0.021	6.968	4.221E-12	0.144	0.021	6.882	7.658E-12
CAGE								
RMhепG2	0.108	0.021	5.150	2.839E-07	0.141	0.021	6.745	1.938E-11
DNase								
RMhепG2	0.041	0.021	1.940	5.253E-02	0.010	0.021	0.467	6.402E-01
H2A.Z								
RMhепG2	0.020	0.021	0.955	3.396E-01	0.008	0.021	0.391	6.959E-01
H3K27ac								
RMhепG2	0.009	0.021	0.411	6.814E-01	-0.011	0.021	-0.502	6.155E-01
H3K27me3								
RMhепG2	-0.015	0.021	-0.727	4.676E-01	0.004	0.021	0.171	8.645E-01
H3K36me3								
RMhепG2	-0.199	0.021	-9.575	2.605E-21	-0.239	0.021	-11.645	1.802E-30
H3K4me1								
RMhепG2	-0.057	0.021	-2.676	7.495E-03	-0.063	0.021	-2.997	2.758E-03
H3K4me2								
RMhепG2	0.081	0.021	3.832	1.306E-04	0.089	0.021	4.234	2.384E-05
H3K4me3								
RMhепG2	-0.030	0.021	-1.434	1.517E-01	-0.029	0.021	-1.384	1.663E-01
H3K79me2								
RMhепG2	0.066	0.021	3.149	1.662E-03	0.055	0.021	2.588	9.727E-03
H3K9ac								
RMhепG2	0.008	0.021	0.360	7.191E-01	0.000	0.021	-0.017	9.863E-01
H3K9me3								
RMhепG2	-0.046	0.021	-2.191	2.856E-02	-0.046	0.021	-2.198	2.803E-02
H4K20me1								
AveRMhепG2	0.098	0.021	4.637	3.739E-06	0.084	0.021	3.977	7.193E-05
CAGE								
AveRMhепG2	0.174	0.021	8.364	1.052E-16	0.226	0.021	10.966	2.733E-27
DNase								
AveRMhепG2	0.054	0.021	2.550	1.083E-02	0.023	0.021	1.091	2.754E-01
H2A.Z								
AveRMhепG2	0.081	0.021	3.843	1.251E-04	0.106	0.021	5.054	4.682E-07
H3K27ac								
AveRMhепG2	0.011	0.021	0.522	6.016E-01	-0.014	0.021	-0.655	5.127E-01
H3K27me3								
AveRMhепG2	-0.012	0.021	-0.562	5.742E-01	0.019	0.021	0.899	3.690E-01
H3K36me3								
AveRMhепG2	-0.171	0.021	-8.203	3.912E-16	-0.203	0.021	-9.809	2.880E-22
H3K4me1								
AveRMhепG2	0.028	0.021	1.329	1.841E-01	0.026	0.021	1.239	2.154E-01
H3K4me2								
AveRMhепG2	0.163	0.021	7.826	7.730E-15	0.174	0.021	8.336	1.327E-16
H3K4me3								
AveRMhепG2	-0.016	0.021	-0.747	4.549E-01	-0.024	0.021	-1.153	2.490E-01
H3K79me2								
AveRMhепG2	0.126	0.021	6.022	2.003E-09	0.144	0.021	6.896	6.921E-12
H3K9ac								

Annotation feature	MT				WT			
	Coeff.	Std. Err.	t value	Pr(> t)	Coeff.	Std. Err.	t value	Pr(> t)
AveRMhепG2 H3K9me3	0.009	0.021	0.432	6.656E-01	0.000	0.021	-0.007	9.947E-01
AveRMhепG2 H4K20me1	-0.032	0.021	-1.528	1.266E-01	-0.033	0.021	-1.538	1.243E-01
MaxRMhепG2 CAGE	0.118	0.021	5.608	2.300E-08	0.099	0.021	4.715	2.566E-06
MaxRMhепG2 DNase	0.171	0.021	8.195	4.157E-16	0.221	0.021	10.714	3.734E-26
MaxRMhепG2 H2A.Z	0.090	0.021	4.276	1.986E-05	0.047	0.021	2.239	2.525E-02
MaxRMhепG2 H3K27ac	0.100	0.021	4.772	1.942E-06	0.137	0.021	6.548	7.235E-11
MaxRMhепG2 H3K27me3	0.039	0.021	1.859	6.318E-02	0.008	0.021	0.385	7.003E-01
MaxRMhепG2 H3K36me3	-0.024	0.021	-1.126	2.603E-01	0.005	0.021	0.255	7.988E-01
MaxRMhепG2 H3K4me1	-0.142	0.021	-6.781	1.525E-11	-0.164	0.021	-7.838	7.048E-15
MaxRMhепG2 H3K4me2	0.101	0.021	4.793	1.750E-06	0.101	0.021	4.794	1.739E-06
MaxRMhепG2 H3K4me3	0.176	0.021	8.445	5.378E-17	0.192	0.021	9.262	4.572E-20
MaxRMhепG2 H3K79me2	0.019	0.021	0.908	3.639E-01	0.018	0.021	0.867	3.861E-01
MaxRMhепG2 H3K9ac	0.151	0.021	7.221	7.041E-13	0.177	0.021	8.502	3.362E-17
MaxRMhепG2 H3K9me3	0.009	0.021	0.442	6.584E-01	-0.001	0.021	-0.044	9.647E-01
MaxRMhепG2 H4K20me1	-0.025	0.021	-1.175	2.401E-01	-0.028	0.021	-1.336	1.815E-01
AveRMhепG2 Methylation	-0.182	0.021	-8.740	4.494E-18	-0.219	0.021	-10.623	9.478E-26
AveRMhепG2 RNAseq	0.049	0.021	2.330	1.989E-02	0.018	0.021	0.872	3.833E-01
MaxRMhепG2 Methylation	-0.158	0.021	-7.551	6.257E-14	-0.199	0.021	-9.578	2.527E-21
MaxRMhепG2 RNAseq	0.109	0.021	5.165	2.616E-07	0.062	0.021	2.945	3.268E-03
GC	0.133	0.021	6.331	2.937E-10	0.158	0.021	7.581	4.986E-14
CpG	0.213	0.021	10.295	2.569E-24	0.209	0.021	10.100	1.748E-23
priPhCons	0.049	0.021	2.304	2.129E-02	0.082	0.021	3.899	9.957E-05
mamPhCons	0.073	0.021	3.449	5.733E-04	0.111	0.021	5.285	1.383E-07
verPhCons	0.069	0.021	3.248	1.179E-03	0.107	0.021	5.109	3.517E-07
priPhyloP	0.021	0.021	0.978	3.284E-01	0.051	0.021	2.407	1.615E-02
mamPhyloP	0.053	0.021	2.519	1.184E-02	0.080	0.021	3.805	1.455E-04
verPhyloP	0.046	0.021	2.172	2.992E-02	0.059	0.021	2.784	5.407E-03
GerpN	0.057	0.021	2.712	6.746E-03	0.065	0.021	3.078	2.109E-03
GerpS	0.035	0.021	1.637	1.019E-01	0.063	0.021	2.991	2.807E-03
GerpRS	0.016	0.021	0.745	4.564E-01	0.031	0.021	1.472	1.413E-01
bStatistic	-0.010	0.021	-0.496	6.199E-01	-0.009	0.021	-0.406	6.845E-01
cHmmTssA	0.194	0.021	9.352	2.023E-20	0.191	0.021	9.176	9.897E-20
cHmmTssAFlnk	0.041	0.021	1.926	5.423E-02	0.075	0.021	3.542	4.055E-04
cHmmTxFlnk	0.085	0.021	4.030	5.759E-05	0.092	0.021	4.357	1.376E-05
cHmmTx	-0.057	0.021	-2.683	7.348E-03	-0.071	0.021	-3.374	7.539E-04

Annotation feature	MT				WT			
	Coeff.	Std. Err.	t value	Pr(> t)	Coeff.	Std. Err.	t value	Pr(> t)
cHmmTxWk	-0.085	0.021	-4.055	5.175E-05	-0.091	0.021	-4.324	1.602E-05
cHmmEnhG	-0.010	0.021	-0.468	6.395E-01	-0.011	0.021	-0.500	6.175E-01
cHmmEnh	0.000	0.021	-0.011	9.910E-01	0.024	0.021	1.131	2.580E-01
cHmmZnfRpts	-0.004	0.021	-0.201	8.405E-01	0.002	0.021	0.096	9.236E-01
cHmmHet	0.015	0.021	0.721	4.713E-01	0.007	0.021	0.325	7.452E-01
cHmmTssBiv	0.048	0.021	2.279	2.273E-02	0.072	0.021	3.389	7.144E-04
cHmmBivFlnk	-0.003	0.021	-0.122	9.031E-01	0.023	0.021	1.077	2.816E-01
cHmmEnhBiv	0.010	0.021	0.453	6.507E-01	0.041	0.021	1.951	5.116E-02
cHmmReprPC	-0.020	0.021	-0.946	3.444E-01	-0.006	0.021	-0.302	7.626E-01
cHmmReprPCWk	-0.032	0.021	-1.514	1.301E-01	-0.022	0.021	-1.027	3.043E-01
cHmmQuies	-0.073	0.021	-3.443	5.854E-04	-0.094	0.021	-4.459	8.641E-06
tOverlapMotifs	0.142	0.021	6.773	1.604E-11	0.206	0.021	9.929	9.159E-23
motifECount	0.121	0.021	5.748	1.028E-08	0.168	0.021	8.055	1.275E-15
motifEHIPos	0.106	0.021	5.020	5.584E-07	0.149	0.021	7.101	1.660E-12
TFBS	0.268	0.020	13.152	4.147E-38	0.297	0.020	14.705	8.545E-47
TFBSPeaks	0.267	0.020	13.108	7.102E-38	0.278	0.020	13.687	5.212E-41
TFBSPeaksMax	0.201	0.021	9.721	6.598E-22	0.240	0.021	11.660	1.535E-30
minDistTSS	-0.006	0.021	-0.301	7.637E-01	-0.006	0.021	-0.300	7.640E-01
CADD	0.116	0.021	5.531	3.561E-08	0.137	0.021	6.522	8.571E-11
MaxGC	0.157	0.021	7.529	7.386E-14	0.212	0.021	10.266	3.418E-24
MaxCpG	0.216	0.021	10.467	4.621E-25	0.222	0.021	10.770	2.100E-26
MaxPriPhCons	0.052	0.021	2.442	1.469E-02	0.084	0.021	3.985	6.967E-05
MaxMamPhCons	0.088	0.021	4.179	3.044E-05	0.121	0.021	5.758	9.678E-09
MaxVerPhCons	0.082	0.021	3.873	1.107E-04	0.115	0.021	5.477	4.802E-08
MaxPriPhyloP	-0.040	0.021	-1.885	5.953E-02	-0.040	0.021	-1.871	6.143E-02
MaxMamPhyloP	0.010	0.021	0.456	6.483E-01	0.020	0.021	0.958	3.381E-01
MaxVerPhyloP	0.037	0.021	1.736	8.267E-02	0.030	0.021	1.434	1.517E-01
MaxGerpN	0.010	0.021	0.494	6.214E-01	0.018	0.021	0.839	4.015E-01
MaxGerpS	0.016	0.021	0.772	4.404E-01	0.023	0.021	1.103	2.703E-01
MaxBStatistic	-0.049	0.021	-2.318	2.053E-02	-0.051	0.021	-2.412	1.593E-02
MaxcHmmTssA	0.205	0.021	9.879	1.476E-22	0.209	0.021	10.083	2.073E-23
MaxcHmm	0.054	0.021	2.564	1.041E-02	0.088	0.021	4.167	3.201E-05
TssAFlnk	0.103	0.021	4.891	1.076E-06	0.105	0.021	4.986	6.638E-07
MaxcHmmTxFlnk	-0.068	0.021	-3.227	1.269E-03	-0.086	0.021	-4.092	4.429E-05
MaxcHmmTx	-0.100	0.021	-4.743	2.239E-06	-0.108	0.021	-5.155	2.760E-07
MaxcHmmTxWk	-0.100	0.021	-4.743	2.239E-06	-0.108	0.021	-5.155	2.760E-07
MaxcHmmEnhG	-0.014	0.021	-0.642	5.212E-01	-0.016	0.021	-0.769	4.419E-01
MaxcHmmEnh	-0.010	0.021	-0.489	6.248E-01	0.015	0.021	0.686	4.926E-01
MaxcHmmZnfRpts	-0.002	0.021	-0.109	9.129E-01	0.006	0.021	0.268	7.888E-01
MaxcHmmHet	0.018	0.021	0.865	3.870E-01	0.010	0.021	0.459	6.462E-01
MaxcHmmTssBiv	0.054	0.021	2.535	1.131E-02	0.075	0.021	3.537	4.126E-04
MaxcHmmBivFlnk	0.008	0.021	0.380	7.037E-01	0.031	0.021	1.447	1.479E-01
MaxcHmmEnhBiv	0.019	0.021	0.880	3.788E-01	0.052	0.021	2.446	1.451E-02
MaxcHmm	-0.017	0.021	-0.799	4.246E-01	-0.005	0.021	-0.258	7.964E-01
ReprPC	-0.017	0.021	-0.799	4.246E-01	-0.005	0.021	-0.258	7.964E-01
MaxcHmm	-0.029	0.021	-1.394	1.635E-01	-0.017	0.021	-0.781	4.346E-01
ReprPCWk	-0.029	0.021	-1.394	1.635E-01	-0.017	0.021	-0.781	4.346E-01
MaxcHmmQuies	-0.081	0.021	-3.855	1.188E-04	-0.103	0.021	-4.892	1.070E-06
MaxTOverlapMotifs	0.122	0.021	5.823	6.599E-09	0.174	0.021	8.327	1.426E-16
MaxMotifECount	0.122	0.021	5.787	8.156E-09	0.186	0.021	8.928	8.846E-19
MaxTFBS	0.289	0.020	14.245	3.873E-44	0.335	0.020	16.808	8.600E-60
MaxTFBSPeaks	0.290	0.020	14.326	1.345E-44	0.317	0.020	15.791	2.481E-53
MaxTFBS	0.204	0.021	9.824	2.487E-22	0.256	0.020	12.519	8.356E-35
PeaksMax	0.204	0.021	9.824	2.487E-22	0.256	0.020	12.519	8.356E-35

Annotation feature	MT				WT			
	Coeff.	Std. Err.	t value	Pr(> t)	Coeff.	Std. Err.	t value	Pr(> t)
MaxMinDistTSS	-0.002	0.021	-0.094	9.252E-01	-0.003	0.021	-0.131	8.962E-01
MaxCADD	0.101	0.021	4.820	1.529E-06	0.096	0.021	4.544	5.819E-06
modARID3A	0.092	0.021	4.387	1.200E-05	0.165	0.021	7.888	4.746E-15
modATF3	0.217	0.021	10.517	2.768E-25	0.296	0.020	14.668	1.415E-46
modBHLHE40	0.121	0.021	5.756	9.772E-09	0.194	0.021	9.330	2.461E-20
modBRCA1	0.164	0.021	7.840	6.918E-15	0.185	0.021	8.905	1.081E-18
modCBX1	0.136	0.021	6.490	1.054E-10	0.172	0.021	8.247	2.742E-16
modCEBPB	0.013	0.021	0.606	5.446E-01	0.095	0.021	4.511	6.793E-06
modCEBPD	0.079	0.021	3.748	1.827E-04	0.162	0.021	7.780	1.102E-14
modCHD2	0.145	0.021	6.934	5.320E-12	0.207	0.021	10.011	4.148E-23
modCTCF	0.046	0.021	2.183	2.915E-02	0.066	0.021	3.107	1.916E-03
modELF1	0.197	0.021	9.485	5.970E-21	0.248	0.020	12.083	1.316E-32
modEP300	0.093	0.021	4.436	9.590E-06	0.154	0.021	7.380	2.216E-13
modEZH2	-0.028	0.021	-1.336	1.816E-01	-0.017	0.021	-0.821	4.116E-01
modFOSL2	0.246	0.021	12.002	3.312E-32	0.326	0.020	16.294	1.725E-56
modFOXA1	-0.006	0.021	-0.263	7.926E-01	0.033	0.021	1.575	1.155E-01
modFOXA2	-0.042	0.021	-1.965	4.956E-02	-0.004	0.021	-0.174	8.619E-01
modFOXK2	0.072	0.021	3.430	6.146E-04	0.109	0.021	5.190	2.290E-07
modGABPA	0.235	0.021	11.436	1.788E-29	0.260	0.020	12.721	7.651E-36
modGATA4	0.084	0.021	3.966	7.534E-05	0.137	0.021	6.539	7.658E-11
modHCFC1	0.205	0.021	9.890	1.325E-22	0.227	0.021	11.008	1.762E-27
modHDAC2	0.048	0.021	2.288	2.225E-02	0.110	0.021	5.221	1.946E-07
modHNF4A	0.018	0.021	0.870	3.845E-01	0.060	0.021	2.846	4.472E-03
modHNF4G	0.004	0.021	0.184	8.542E-01	0.047	0.021	2.232	2.569E-02
modIRF3	0.106	0.021	5.052	4.725E-07	0.146	0.021	6.956	4.589E-12
modJUN	0.191	0.021	9.214	7.055E-20	0.287	0.020	14.183	8.792E-44
modJUND	0.254	0.020	12.426	2.495E-34	0.357	0.020	18.037	5.273E-68
modMAFF	0.127	0.021	6.045	1.746E-09	0.160	0.021	7.676	2.435E-14
modMAFK	0.120	0.021	5.708	1.297E-08	0.144	0.021	6.857	9.063E-12
modMAX	0.147	0.021	7.032	2.692E-12	0.188	0.021	9.049	3.055E-19
modMAZ	0.126	0.021	5.996	2.353E-09	0.206	0.021	9.945	7.818E-23
modMBD4	0.049	0.021	2.315	2.069E-02	0.088	0.021	4.176	3.087E-05
modMXI1	0.154	0.021	7.386	2.123E-13	0.199	0.021	9.585	2.373E-21
modMYBL2	0.119	0.021	5.683	1.499E-08	0.172	0.021	8.247	2.747E-16
modMYC	0.121	0.021	5.740	1.077E-08	0.183	0.021	8.785	3.054E-18
modNFIC	0.079	0.021	3.726	1.994E-04	0.129	0.021	6.127	1.052E-09
modNR2C2	0.132	0.021	6.285	3.932E-10	0.167	0.021	8.026	1.605E-15
modNRF1	0.073	0.021	3.453	5.654E-04	0.107	0.021	5.067	4.375E-07
modPOLR2A	0.091	0.021	4.328	1.573E-05	0.127	0.021	6.047	1.729E-09
modPOLR2A phosphoS2	0.094	0.021	4.446	9.160E-06	0.127	0.021	6.072	1.484E-09
modPOLR2A phosphoS5	0.095	0.021	4.493	7.386E-06	0.122	0.021	5.822	6.659E-09
modRAD21	0.016	0.021	0.779	4.361E-01	0.046	0.021	2.176	2.968E-02
modRCOR1	0.098	0.021	4.650	3.517E-06	0.139	0.021	6.648	3.732E-11
modREST	0.138	0.021	6.584	5.676E-11	0.187	0.021	8.995	4.904E-19
modRFX5	0.160	0.021	7.640	3.202E-14	0.202	0.021	9.748	5.114E-22
modRXRA	0.036	0.021	1.689	9.128E-02	0.092	0.021	4.362	1.350E-05
modSIN3A	0.135	0.021	6.431	1.548E-10	0.165	0.021	7.887	4.795E-15
modSIN3B	0.167	0.021	7.992	2.115E-15	0.164	0.021	7.873	5.345E-15
modSMC3	0.006	0.021	0.302	7.627E-01	0.053	0.021	2.510	1.215E-02
modSP1	0.070	0.021	3.307	9.593E-04	0.127	0.021	6.034	1.870E-09
modSP2	0.181	0.021	8.675	7.827E-18	0.208	0.021	10.040	3.149E-23
modSRF	0.145	0.021	6.908	6.364E-12	0.181	0.021	8.688	7.013E-18

Annotation feature	MT				WT			
	Coeff.	Std. Err.	t value	Pr(> t)	Coeff.	Std. Err.	t value	Pr(> t)
modTAF1	0.153	0.021	7.340	2.978E-13	0.190	0.021	9.125	1.553E-19
modTBP	0.138	0.021	6.586	5.626E-11	0.188	0.021	9.065	2.661E-19
modTCF12	0.029	0.021	1.370	1.707E-01	0.098	0.021	4.671	3.171E-06
modTCF7L2	-0.018	0.021	-0.846	3.975E-01	-0.001	0.021	-0.040	9.682E-01
modTEAD4	0.031	0.021	1.458	1.451E-01	0.087	0.021	4.144	3.541E-05
modTFAP4	0.087	0.021	4.147	3.494E-05	0.138	0.021	6.573	6.099E-11
modUSF1	0.114	0.021	5.423	6.509E-08	0.171	0.021	8.214	3.566E-16
modUSF2	0.120	0.021	5.733	1.121E-08	0.172	0.021	8.229	3.179E-16
modYY1	0.137	0.021	6.541	7.532E-11	0.168	0.021	8.049	1.338E-15
modZBTB33	0.028	0.021	1.323	1.861E-01	0.052	0.021	2.473	1.347E-02
modZBTB7A	0.028	0.021	1.301	1.933E-01	0.109	0.021	5.167	2.587E-07
modZHX2	0.176	0.021	8.428	6.197E-17	0.223	0.021	10.814	1.336E-26
modZKSCAN1	0.116	0.021	5.504	4.134E-08	0.151	0.021	7.244	5.977E-13
modZNF274	0.001	0.021	0.065	9.482E-01	0.053	0.021	2.532	1.140E-02
OCChIPOnly	0.026	0.021	1.249	2.119E-01	0.016	0.021	0.753	4.516E-01
OCDNaseOnly	0.002	0.021	0.117	9.066E-01	0.000	0.021	-0.010	9.924E-01
OCFAIREOnly	-0.048	0.021	-2.292	2.202E-02	-0.095	0.021	-4.498	7.220E-06
OCOpenChrom	0.034	0.021	1.607	1.082E-01	0.017	0.021	0.799	4.243E-01
OCUnDef	-0.099	0.021	-4.713	2.592E-06	-0.128	0.021	-6.095	1.286E-09
OCValid	0.072	0.021	3.419	6.394E-04	0.133	0.021	6.350	2.599E-10
ChromHMM1	0.201	0.021	9.696	8.414E-22	0.205	0.021	9.883	1.418E-22
ChromHMM2	-0.086	0.021	-4.096	4.346E-05	-0.080	0.021	-3.776	1.638E-04
ChromHMM3	-0.054	0.021	-2.534	1.135E-02	-0.066	0.021	-3.141	1.704E-03
ChromHMM4	-0.053	0.021	-2.529	1.149E-02	-0.049	0.021	-2.325	2.015E-02
ChromHMM5	-0.023	0.021	-1.079	2.807E-01	-0.019	0.021	-0.900	3.682E-01
ChromHMM6	-0.007	0.021	-0.334	7.382E-01	0.008	0.021	0.358	7.207E-01
ChromHMMUnDef	-0.021	0.021	-0.994	3.203E-01	-0.028	0.021	-1.327	1.845E-01
SegWayBRD	-0.001	0.021	-0.063	9.494E-01	-0.034	0.021	-1.604	1.089E-01
SegWayCON	0.001	0.021	0.047	9.622E-01	-0.001	0.021	-0.064	9.492E-01
SegWayFAC	0.022	0.021	1.057	2.906E-01	0.034	0.021	1.596	1.107E-01
SegWayQUI	0.012	0.021	0.584	5.594E-01	0.020	0.021	0.968	3.330E-01
SegWaySPC	-0.016	0.021	-0.745	4.565E-01	0.005	0.021	0.244	8.070E-01
SegWayUnDef	-0.007	0.021	-0.329	7.422E-01	-0.004	0.021	-0.175	8.610E-01
SegwayC0	-0.006	0.021	-0.266	7.905E-01	0.015	0.021	0.729	4.663E-01
SegwayC1	-0.039	0.021	-1.865	6.229E-02	-0.038	0.021	-1.810	7.050E-02
SegwayD	-0.045	0.021	-2.132	3.311E-02	-0.046	0.021	-2.153	3.141E-02
SegwayE/GM	-0.003	0.021	-0.127	8.987E-01	-0.023	0.021	-1.079	2.809E-01
SegwayF0	-0.015	0.021	-0.687	4.919E-01	0.003	0.021	0.142	8.873E-01
SegwayF1	-0.033	0.021	-1.554	1.202E-01	-0.029	0.021	-1.366	1.722E-01
SegwayGE0	0.011	0.021	0.507	6.124E-01	-0.024	0.021	-1.116	2.646E-01
SegwayGE1	-0.037	0.021	-1.762	7.816E-02	-0.061	0.021	-2.882	3.993E-03
SegwayGE2	-0.044	0.021	-2.102	3.570E-02	-0.009	0.021	-0.410	6.820E-01
SegwayGM0	0.018	0.021	0.830	4.067E-01	0.020	0.021	0.932	3.512E-01
SegwayGM1	-0.016	0.021	-0.777	4.373E-01	-0.031	0.021	-1.448	1.476E-01
SegwayGS	0.031	0.021	1.474	1.406E-01	0.065	0.021	3.078	2.106E-03
SegwayH3K9me1	-0.041	0.021	-1.953	5.097E-02	-0.041	0.021	-1.918	5.524E-02
SegwayL0	0.004	0.021	0.203	8.391E-01	0.021	0.021	0.985	3.246E-01
SegwayL1	-0.046	0.021	-2.189	2.870E-02	-0.043	0.021	-2.055	3.998E-02
SegwayR0	-0.039	0.021	-1.867	6.197E-02	-0.029	0.021	-1.373	1.699E-01
SegwayR1	-0.017	0.021	-0.804	4.214E-01	-0.037	0.021	-1.731	8.360E-02
SegwayR2	-0.043	0.021	-2.027	4.277E-02	-0.052	0.021	-2.450	1.436E-02
SegwayR3	-0.034	0.021	-1.622	1.049E-01	-0.041	0.021	-1.955	5.073E-02
SegwayR4	-0.030	0.021	-1.417	1.566E-01	-0.059	0.021	-2.797	5.194E-03
SegwayR5	0.006	0.021	0.307	7.590E-01	-0.012	0.021	-0.549	5.832E-01

Annotation feature	MT				WT			
	Coeff.	Std. Err.	t value	Pr(> t)	Coeff.	Std. Err.	t value	Pr(> t)
SegwayTF0	-0.008	0.021	-0.378	7.056E-01	0.009	0.021	0.436	6.628E-01
SegwayTF1	-0.019	0.021	-0.890	3.733E-01	-0.013	0.021	-0.625	5.319E-01
SegwayTF2	-0.019	0.021	-0.913	3.614E-01	-0.001	0.021	-0.049	9.607E-01
SegwayTSS	0.222	0.021	10.744	2.742E-26	0.239	0.021	11.639	1.927E-30
SegwayUnDef	-0.012	0.021	-0.559	5.765E-01	0.000	0.021	-0.010	9.917E-01
ERBFeatureCTCF	0.015	0.021	0.690	4.900E-01	0.020	0.021	0.953	3.408E-01
ERBFeatureenhancer	-0.037	0.021	-1.732	8.336E-02	-0.048	0.021	-2.263	2.372E-02
ERBFeatureopen_ chromatin	-0.047	0.021	-2.201	2.784E-02	-0.047	0.021	-2.228	2.601E-02
ERBFeaturepromoter	0.165	0.021	7.930	3.442E-15	0.177	0.021	8.516	2.993E-17
ERBFeaturepromoter _flanking	-0.045	0.021	-2.141	3.240E-02	-0.027	0.021	-1.292	1.965E-01
ERBFeatureTF	-0.017	0.021	-0.795	4.266E-01	-0.027	0.021	-1.281	2.004E-01
ERBFeatureUnDef	-0.050	0.021	-2.367	1.802E-02	-0.068	0.021	-3.206	1.365E-03

1320

1321

Additional files

1322

1323

1324 *File 1:* Annotated plasmid sequence file in GenBank Flat File Format.

1325

1326

1327 *File 2:* Designed array oligo nucleotide sequences (gzip-compressed text file)

1328

1329

1330 *File 3:* Text file listing out annotations used for prediction of observed RNA/DNA ratios.

1331